Genetic Association and Prediction Methods for Biobank Data

by

Zhangchen Zhao

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2021

Doctoral Committee:

Associate Professor Seunggeun Lee, Co-Chair
Professor Bhramar Mukherjee, Co-Chair
Associate Research Scientist Lars Fritsche
Associate Professor Jennifer Smith

Zhangchen Zhao

zczhao@umich.edu

ORCID iD: 0000-0002-7990-2589

To my family for their love and support!

ACKNOWLEDGEMENTS

Since I was a little kid, it has been my dream to pursue a doctorate degree. However, the road towards this goal was never easy. I met countless difficulties and sometimes felt frustrated. Luckily, I have not been alone. This dissertation would not have been possible without the support and guidance from my committee, friends, and family.

First, I would like to express my deepest gratitude to my thesis advisor, Dr. Seunggeun Shawn Lee. Shawn introduced me to statistical genetics step by step and provided me invaluable guidance of research experiences, which helped me build a solid foundation of statistical genetics. Moreover, he has always given me freedom to pursue different research directions, provided insightful suggestions, asked questions to help me think deeper, and encouraged me to overcome challenges. He also always answered all my questions patiently, no matter how silly they were. His extraordinary mentorship helped me become a better student and also a better researcher.

I would like to extend my gratitude to my co-advisor Dr. Bhramar Mukherjee. Besides dissertation work, I also worked on four side projects with her. With her guidance and help, I successfully published my first paper in 2017. Not only did she offer me her guidance on research but her dedication and enthusiasm in science has profoundly inspired me how to be a good scholar.

I am grateful for my committee member Dr. Lars Fritsche for helping me have a deeper understanding of genetic data and working on the serology data analysis with me. It is really my honor and pleasure to have the chance collaborating with him on up to five papers.

I am also very appreciative to Dr. Cristen Willer. She allowed me to attend her lab meeting and I learnt more genetic knowledge from the view of a bioinformatician and geneticist.

I would also like express my heartfelt thanks to Dr. Jennifer Smith for kindly joining my committee when Cristen cannot attend my dissertation defense. I really appreciate her valuable and insightful advice for my research projects.

Additionally, I would like to express my sincere thanks to Kirsten Herold, who has offered tremendous help on my paper writing. The same gratitude goes to the members of Shawn's lab for their great collaboration and discussion.

Last but not least, I would like to thank my wife Hongyou Chen, my parents, my parents-in-law, and my grandparents for their continuous encouragement and support. They always cheer me up, listen to my fears, and share my happiness. It is their love that helps me get through all the hard times and keep me fearless to face challenges. It is very lucky for me to have met so many great people, and I feel honored to have them always accompany me either in person or virtually.

TABLE OF CONTENTS

DEDICATION ..... **ii**

ACKNOWLEDGEMENTS ..... **iii**

LIST OF FIGURES ..... **viii**

LIST OF TABLES ..... **xii**

LIST OF ABBREVIATIONS ..... **xv**

ABSTRACT ..... **xvii**

CHAPTER I: Introduction **1**

1.1 Motivation 1

1.2 Challenges of analyzing biobank data 2

1.2.1 Unbalanced case-control ratios 2

1.2.2 Sample relatedness 2

1.2.3 Translation to minor ancestry group 3

1.3 Summary of objectives 4

CHAPTER II: UK-Biobank Whole Exome Sequence Binary Phenome Analysis with Robust Region-based Rare Variant Test **5**

2.1 Introduction 7

2.2 Methods 9

2.2.1 Gene/region-based rare variant tests for binary traits 9

2.2.2 Saddle Point Approximation (SPA) and Efficient Resampling (ER) 10

2.2.3 Robust burden test, robust SKAT and robust SKAT-O 11

2.2.4 Extension to the joint test of common and rare variants 12

2.2.5 Numerical Simulations 13

2.2.6 Analysis of whole exome sequencing (WES) data in the UK Biobank 14

2.3 Results 16

2.3.1 Type I Error and Power Simulation Results 16

2.3.2 Comparison of computational times 19

2.3.3 Analysis of whole exome sequencing (WES) data in the UK Biobank 20
2.4 Discussion 24
2.5 Supplementary Materials.....28
2.5.1 Supplementary Figures 28
2.5.2 Supplementary Tables.....37
CHAPTER III: Scalable Generalized Linear Mixed Model for Region-based Association Tests in Large Biobanks and Cohorts.....47
3.1 Introduction 48
3.2 Methods.....50
3.2.1 Overview of methods.....50
3.2.2 Generalized linear mixed model 53
3.2.3 Estimate variance component and other model parameters (Step 1).....53
3.2.4 Gene-based association tests (Step 2).....54
3.2.5 Approximate $\tilde{G}^T \hat{P} \tilde{G}$ 55
3.2.6 Robust adjustment for $\hat{R}_S^{\frac{1}{2}} \tilde{G}^T \hat{P}_S \tilde{G} \hat{R}_S^{\frac{1}{2}}$ to account for unbalanced case-control ratios..56
3.2.7 Conditional analysis 57
3.2.8 Data simulation 58
3.2.9 HUNT and UK Biobank data analysis 60
3.3 Results.....61
3.3.1 Computation and Memory Cost 61
3.3.2 Gene-based association analysis of quantitative traits in HUNT and UK Biobank .63
3.3.3 Gene-based association analysis of binary traits in UK Biobank 68
3.3.4 Simulation Studies 68
3.4 Discussion 71
3.5 Supplementary Materials 75
3.5.1 Algorithm details 75
3.5.2 Additional simulation and real-data analysis results 83
3.5.3 Supplementary figures 88
3.5.4 Supplementary tables 99

CHAPTER IV: The Construction of Multi-ethnic Polygenic Risk Score Using Transfer Learning **123**

4.1 Introduction 125

4.2 Methods 127

4.2.1 Polygenic risk score using GWAS summary statistics from a single ancestry. 127

4.2.2 Transfer learning (TL-PRS) using GWAS summary statistics from a single ancestry. 128

4.2.3 Combining multiple GWAS summary statistics from different ancestries. 129

4.2.4 Simulations using South Asian samples in UK Biobank 130

4.2.5 Analysis of South Asian, African, and non-British White samples in UK Biobank 131

4.3 Results 132

4.3.1 Overview of TL-PRS 132

4.3.2 Simulations using South Asian samples in UK Biobank 134

4.3.3 Prediction performance of South Asian, African, and non-British White samples in UK Biobank 138

4.4 Discussion 142

4.5 Supplementary Materials 146

4.5.1 Supplementary Figures 146

4.5.2 Supplementary Tables 150

CHAPTER V: Summary and Discussion **175**

5.1 Summary 175

5.2 Extension and future work 177

BIBLIOGRAPHY **180**

LIST OF FIGURES

Figure

2.1 Empirical power estimates for the unadjusted and robust versions of SKAT-O, and hybrid method 18

2.2 Comparison of computation time of unadjusted, hybrid, ER and robust approaches for SKAT-O 20

2.3 PheWAS plots of 10 rare variant associations with p-value< 10^{-7} 23

S2.1 Empirical power estimates for robust SKAT, burden, SKAT-O with the same number of cases across different case control ratios 28

S2.2 The distribution of the number of variants in genes in the UK-Biobank WES data 30

S2.3 Empirical power estimates for the unadjusted and robust version of SKAT and burden test where 30% of variants were causal variants and all causal variants were risk-increasing 31

S2.4 Empirical power estimates for the unadjusted and robust versions of SKAT-O and hybrid method where 30% of variants were causal variants. 80% causal variants were risk-increasing and 20% were risk-deceasing 32

S2.5 Empirical power estimates for the unadjusted and robust versions of SKAT-O and hybrid method where 30% of variants were causal variants. All causal variants were risk-increasing 33

S2.6 Empirical power estimates for robust SKAT, burden and SKAT-O where 30% of variants were causal variants and all causal variants were risk-increasing 34

S2.7 P-values of single variants in 10 significant genes 35

S2.8 QQ Plots of SKAT-O p-values of 10 selected phenotypes 36

3.1 Estimated and projected computation cost by sample sizes (N) for gene-based tests for 15,342 genes, each containing 50 rare variants 52

3.2 Quantile-quantile plots of exome-wide gene-based association results for (A) high-density lipoprotein (HDL) in the HUNT study (N = 69,214); (B) automated read pulse rate in the UK Biobank (N = 385,365). C. glaucoma in the UK Biobank (N cases = 4,462; N controls = 397,761) 66

S3.1 Workflow of SAIGE-GENE 88

S3.2 Plots of the variance ratio of the score statistics by MAC for rare variants with and without the full GRM for sample relatedness (left) and with the full GRM and a sparse GRM for closely related samples(right) 89

S3.3 Scatter plots of association p-values from SAIGE-GENE versus SMMAT(Chen et al., 2018) and EmmaX-SKAT for the Burden, SKAT, and SKAT-O tests based on simulation data on the -log10 scale 90

S3.4 Scatter plots of association p-values from SAIGE-GENE versus SMMAT and EmmaX-SKAT for the Burden, SKAT, and SKAT-O tests based on real data analysis on the -log10 scale 91

S3.5 Scatter plots of association p-values on the -log10 scale from SAIGE-GENE with two sample relatedness cutoffs for the sparse GRM, 0.125 and 0.2. 15,338 genes were tested for automated read pulse rate in white British samples in the HRC-imputed UK Biobank 92

S3.6 Quantile-quantile plots of association p-values for 10 million variant sets from the simulation study for phenotypes with various case-control ratios 92

S3.7 Empirical computation time for A. step 1 for fitting a null mixed model and B. step 2 for association tests, respectively by sample sizes (N) for gene-based tests for 15,342 genes, each containing 50 rare variants 93

S3.8 Log-log plot of the estimated run time as a function of number of markers per gene 94

S3.9 Log-log plots of the estimated A. run time and B. memory usage as a function of sample size (N) for genome-wide tests for 286,000 chunks, each containing 50 variants on

average, given that there are 14.3 million markers in the HRC-imputed UK Biobank with MAF ≤ 1% and imputation info score ≥ 0.8 95

S3.10 Log-log plots of the estimated run time for as a function of sample size (N) for SAIGE-GENE with and without using the robust adjustment 96

S3.11 Pedigree of families, each with 10 members, in the simulation study 97

S3.12 Histogram of simulated phenotypes 97

S3.13 Comparing heritability estimates using the sparse GRM to heritability estimates using the full GRM for 24 quantitative traits in the UK Biobank with sample size (N) ≥ 100,000 98

4.1 Overview of transfer learning on PRS methods. LD ref: LD reference panel. (A) The general procedure how to construct PRS using transfer learning; (B) The procedure how to combine multiple PRS into the final PRS; (C) The detailed procedure of transfer learning 133

4.2 Relative accuracy of transfer learning method as a function of iterations. (A) The simulation setting where the causal markers were 0.1%, genetic correlation was 0.4 and European summary statistics were used; (B) The real data analysis of HDL in a South Asian cohort from UK Biobank, where UKBB summary statistics were used 134

4.3 Prediction accuracy of single-source and multi-source polygenic prediction methods in simulations. Different genetic designs (0.1% and 1% causal variants) were simulated as well as cross-population genetic correlations (0.4, 0.7 and 1.0). 135

4.4 Relative prediction accuracy of single-source and multi-source polygenic prediction methods using transfer learning, with respect to the base models without transfer learning across 20 replicates in the simulation 137

4.5 Relative prediction accuracy of single-source and multi-source polygenic prediction methods using transfer learning, with respect to the base models without transfer learning across 8 traits in South Asian, African and Non-British White. Each point shows the relative prediction R2 of a trait 141

S4.1 Relative prediction accuracy of single-source and multi-source polygenic prediction methods with respect to lsum trained using UKBB GWAS across 8 traits in the cohorts South Asian, African and Non-British White.....146

S4.2 Cumulant event plot in terms of the top 10% PRS constructed by transfer learning methods and their base methods.....147

LIST OF TABLES

Table

2.1 Type I error rates of unadjusted and robust versions of burden, SKAT and SKAT-O and hybrid method 16

2.2 Significant gene-phenotype associations in the UK Biobank WES data 22

S2.1 Type I error rate divided by α of different methods when testing an association with dichotomous traits at stringent α levels $\alpha = 10^{-2}, 10^{-4}$ and 2.5×10^{-6} 37

S2.2 Type I error rate divided by α of robust SKAT-O when testing an association with dichotomous traits at stringent α levels $\alpha = 10^{-2}, 10^{-4}$ and 2.5×10^{-6} with three different region length (1) 1kb; (2) 2kb; (3) 3kb 38

S2.3 Type I error rate divided by α of different methods when testing an association between all variants, including both common and rare variants, and dichotomous traits at stringent α levels $\alpha = 10^{-2}, 10^{-4}$ and 2.5×10^{-6} 39

S2.4 Gene-phenotype associations detected by Robust SKAT-O across 791 phenotypes at α=2.5×10-6 (Number of associations=111) 40

S2.5 Top 3 single rare-variant signals of associations with p-value < 10-7 in the UK Biobank WES data 44

S2.6 The most significant nearby variant association signals (± 100 Kbp up and down stream) in the UK-Biobank imputed datasets of 400,000 British samples 46

3.1 Genes that are significantly associated with automated read pulse rate and glaucoma in the UK Biobank and high-density lipoprotein (HDL) in the HUNT study with SKAT-O p-values < 2.5x10-6 from SAIGE-GENE 65

S3.1 The estimated run time (A) and memory use (B) across different sample sizes 99

S3.2 The estimated run time (A) and memory use (B) across different sample sizes for binary traits with and without the robust adjustment 102

S3.3 Heritability estimated based on the full GRM by step 1 in SAIGE-GENE A. for 53 quantitative traits and B. for 10 binary traits in the UK Biobank 105

S3.4 Exome-wide significant genes with p-values ≤ 2.5x10-6 identified by SAIGE-GENE in the UK Biobank a. for 53 quantitative traits, b for 10 binary traits with various case-control ratios 107

S3.5 Exome-wide significant genes with p-values ≤ 2.5x10-6 identified by SAIGE-GENE but not identified by SAIGE in the UK Biobank for 53 quantitative traits 111

S3.6 Exome-wide significant genes with p-values ≤ 2.5x10-6 identified by SAIGE-GENE and remained significant after conditioning on the most significant variant, given that the most significant variant is a common variant with MAF > 1% or a less frequent non-coding variant that is not included in the gene-based tests for A. 53 quantitative traits B. 6 binary traits 112

S3.7 Empirical type I error rates for SAIGE-GENE, SAIGE-GENE-GCadj (GC adjusted SAIGE-GENE), EmmaX-SKAT(Kang et al., 2010; M. C. Wu et al., 2011) and SMMAT(Chen et al., 2018). h^2: heritability 117

S3.8 Empirical type I error rates for SAIGE-GENE with the larger sample size of 1,000 families and 10,000 independent samples (total sample size N = 20,000). The heritability $h^2 = 0.2$ 118

S3.9 Empirical type I error rates for SAIGE-GENE, EmmaX-SKAT and SMMAT for skewed distributed phenotypes with and without inverse normal transformation for 500 families and 5,000 independent samples (total sample size N = 10,000) 118

S3.10 Empirical type I error rates for SAIGE-GENE and SMMAT for skewed distributed phenotypes with the three-step phenotype transformation procedure for 500 families and 5,000 independent samples (total sample size N = 10,000) 119

S3.11 Empirical type I error rates for SAIGE-GENE in the presence of population stratification 120

S3.12 Empirical type I error rates for SAIGE-GENE in the presence of non-negligible cryptic relatedness 120

S3.13 Empirical Type I error rates of SAIGE-GENE for binary traits with five different prevalence 121

S3.14 Empirical type I error rates for SAIGE-GENE in the simulation study with case-control sampling from an underlying large cohort 122

S3.15 Empirical power for SAIGE-GENE and EmmaX-SKAT with two different percentages of causal variants (top vs bottom panels) and two different ratios of positive and negative effect directions (left vs right) 122

S3.16 Empirical power for the SKAT-O test in SAIGE-GENE for binary phenotypes in cohort studies and case-control sampling studies 122

4.1 Prediction accuracy of different single-source and multi-source polygenic prediction methods in analyses of LDL in the African cohort of UK Biobank 139

S4.1 List of data sets used in simulations and analyses of real phenotypes 150

S4.2 Prediction accuracy of different PRS construction methods in analyses of eight traits in the South Asian cohort of UK Biobank 151

S4.3 Prediction accuracy of different PRS construction methods in analyses of eight traits in the African cohort of UK Biobank 159

S4.4 Prediction accuracy of different PRS construction methods in analyses of eight traits in the Non-British White cohort of UK Biobank 167

LIST OF ABBREVIATIONS

AFR: African

BBJ: Biobank Japan

BMI: Body Mass Index

CAD: Coronary Artery Disease

DBP: Diastolic Blood Pressure

EHR: Electronic Health Record

ER: Efficient Resampling

GLMM: Generalized Linear Mixed Model

GWAS: Genome-wide Association Studies

HDL: High-density Lipoproteins

LD: Linkage Disequilibrium

LDL: Low-density Lipoproteins

LMM: Linear Mixed Model

LRT: Likelihood Ratio Test

MAC: Minor Allele Count

MAF: Minor Allele Frequency

MGI: Michigan Genomics Initiative

NBW: Non-British White

PC: Principal Component

PheWAS: Phenome-Wide Association Studies

PRS: Polygenic Risk Score

PT: Pruning and thresholding

SAS: South Asian

SBP: Systolic Blood Pressure

SKAT: SNP-Set (Sequence) Kernel Association Test

SNP: Single Nucleotide Polymorphism

SPA: Saddlepoint Approximation

TG: Triglycerides

T2D: Type-2 Diabetes

UKBB: UK Biobank

WES: Whole Exome Sequencing

ABSTRACT

With large sample sizes, population-based cohorts and biobanks provide an exciting opportunity to identify genetic components of complex traits. For example, UK Biobank provides genome-wide genotyping data of 500,000 volunteer participants, which is an invaluable resource to detect genetic associations and build prediction models of genetic effects. In the first two projects, we focus on discovery-type questions and develop robust region-based tests of genetic association. In the third project, we target a translation-type question and develop a multi-ethnic prediction method.

In the first project, we propose SKAT/SKAT-O type region-based tests to account for unbalanced case-control ratios. In biobank data analysis, most binary phenotypes have unbalanced case-control ratios, which can cause inflation of type I error rates. Recently, a saddlepoint approximation (SPA) based single variant test has been developed to provide an accurate and scalable method to test for associations of such phenotypes. For gene- or region-based multiple variant tests, a few methods exist that can adjust for unbalanced case-control ratios; however, these methods are either less accurate when case-control ratios are extremely unbalanced or not scalable for large data analyses. To address these problems, we develop a robust method, where the single-variant score statistic is calibrated based on SPA and Efficient Resampling (ER). Through simulation studies, we show that the proposed method provides well-calibrated p-values. The proposed method has similar computation time as the unadjusted approaches and is scalable for large sample data. In our application, the UK Biobank whole exome sequence data analysis of 45,596 unrelated European

samples and 791 PheCode phenotypes identified 10 rare variant associations with p-value < 10^{-7}, including the associations between *JAK2* and myeloproliferative disease, *HOXB13* and cancer of prostate, and *F11* and congenital coagulation defects.

In the second project, we extend the robust method to related samples. Here we propose a scalable generalized mixed model region-based association test that can handle large sample sizes and accounts for unbalanced case-control ratios for binary traits. This method, SAIGE-GENE, utilizes state-of-the-art optimization strategies to reduce computational and memory cost, and hence is applicable to exome-wide and genome-wide region-based analysis for hundreds of thousands of samples. Through the analysis of the HUNT study of 69,716 Norwegian samples and the UK Biobank data of 408,910 White British samples, we show that SAIGE-GENE can efficiently analyze large sample data (N > 400,000) with type I error rates well controlled.

In the third project, we propose a novel multi-ethnic PRS using transfer learning from machine learning literature. As most existing GWAS were conducted in European or East Asian individuals, the existing PRS models have limited transferability to minority populations such as Africans and South Asians. Although recent studies have developed multi-ethnic PRS models that linearly combine multiple PRS trained with different ancestry GWAS, they remain under-powered. Our approach, TL-PRS, fine-tunes the potentially biased model trained with GWAS summary statistics from the majority ancestry to the target dataset of the minority ancestry. Through simulation studies, we show that TL-PRS improved the performance of PRS with a wide range of genetic architectures and cross-population genetic correlations compared to the baseline methods. In the application of 8,168 Africans and 10,285 South Asians of UK Biobank data, TL-PRS substantially improved the prediction accuracy of the six quantitative and two dichotomous traits.

CHAPTER I

Introduction

1.1 Motivation

With large sample sizes, population-based cohorts and biobanks provide an exciting opportunity to identify genetic components of complex traits. For example, UK Biobank is an international health resource following the health and well-being of 500,000 volunteer participants, whose genome-wide genotyping data are all available.(Sudlow et al., 2015) In addition to the availability of genetics data, the wealth of information of electronic health records (EHRs) can be leveraged to improve our understanding of the genetic relationship with human diseases by describing a landscape of genetic associations across many different measures, such as disease diagnosis codes.(Pendergrass & Ritchie, 2015)

With the availability of genetic data and EHRs, phenome-wide association studies (PheWASs) have emerged as an important tool for identifying comprehensive genetic associations between SNPs and a wide range of phenotypes, with successful implementation in phenotype data collected from EHRs. We are motivated by these abundant data resources to work on genetic association and prediction methods for biobank data.

Recently, rare-variant analysis has become increasingly important. In PheWAS analysis, it was initially assumed that common genetic variants made a main contribution to common human

diseases, such as cardiovascular diseases. However, the proportion of variation explained by common variants seems very modest; moreover, there are very few examples of the actual variant being identified. Instead, additional disease risk can be explained by rare functional variants, which provides further challenges for genetic studies.

1.2 Challenges of analyzing biobank data

A biobank is a repository that stores human genetic and biological data. Since the late 1990s, biobanks have become an important resource supporting many types of contemporary research, such as genetic studies. However, there are several well-known challenges when analyzing biobank data, such as huge data size and ethical problems surrounding privacy. In this dissertation, we mainly focus on the following three problems.

1.2.1 Unbalanced case-control ratios

Because most diseases and symptoms have low prevalence in general, the binary phenotypes in biobanks usually have unbalanced or extremely unbalanced case-control ratios, such as 1:10 or 1:100. For example, in the UK Biobank data, nearly 99% of PheCode-based binary phenotypes have case-control ratios of less than 1:10. Substantial challenges are posed when analyzing the associations between rare variants and unbalanced phenotypes. For gene- or region-based multiple variant tests, a few methods exist that can adjust for unbalanced case-control ratios; however, these methods are either less accurate when case-control ratios are extremely unbalanced or not scalable for large data analyses.

1.2.2 Sample relatedness

Sample relatedness and population stratification are major confounders in genetic association studies and need to be controlled in PheWAS analysis. Sample relatedness may violate the independence assumption required by some models, such as linear regression. Linear mixed models (LMM) are widely used to account for these issues in GWAS for both binary and quantitative traits. However, since LMM is not designed to analyze binary traits, it can have inflated type I error rates, especially in the presence of unbalanced case-control ratios.

1.2.3 Translation to minor ancestry group

Although UK Biobank reveals hundreds of thousands of European genetic data, genetic data of other populations are still limited for analysis, such as South Asians and Africans. Due to inadequate sample sizes, it is almost impossible to identify the genetic associations within these populations. Hence, it is an obvious idea to utilize European summary results on other populations to further improve the prediction accuracy of disease risks. However, due to different patterns of linkage disequilibrium (LD) and potentially different causal effects, the use of European data for polygenic risk prediction in non-European populations reduces prediction accuracy.(Márquez-Luna et al., 2017; Purcell et al., 2009; Vilhjálmsson et al., 2015) For example, Vilhjálmsson et al. (2015) reported that Japanese and African-American populations had a relative decrease of 53–89% in schizophrenia risk prediction accuracy when applying prediction methods using European data.(Vilhjálmsson et al., 2015) Although recent studies have developed multi-ethnic PRS models(Márquez-Luna et al., 2017) that linearly combine multiple PRS trained with different ancestry GWAS, they remain under-powered.

The proposed methods in the following chapters are going to address these three challenges.

1.3 Summary of Objectives

With a focus on the challenges presented above, in this dissertation, I present methodologies that are aimed to achieve the following analytical objectives:

1) To develop robust region-based methods for unbalanced case-control ratios within independent samples to control type I error rates;
2) To extend the robust methods to generalized linear mixed model (GLMM) to adjust for relatedness among samples;
3) To build multi-ethnic polygenic risk score (PRS) models to improve the polygenic risk prediction in non-European individuals using summary statistics from UK Biobank (UKBB) and Biobank Japan (BBJ).

The three objectives are addressed in the proposed methods in Chapter II, Chapter III, and Chapter IV respectively. More details on the background, pertinent literature, motivation and methodology development, can be found in the introduction sections of each chapter.

CHAPTER II

UK-Biobank Whole Exome Sequence Binary Phenome Analysis with Robust Region-based Rare Variant Test

Abstract

In biobank data analysis, most binary phenotypes have unbalanced case-control ratios, which can cause inflation of type I error rates. Recently, a saddlepoint approximation (SPA) based single variant test has been developed to provide an accurate and scalable method to test for associations of such phenotypes. For gene- or region-based multiple variant tests, a few methods exist that can adjust for unbalanced case-control ratios; however, these methods are either less accurate when case-control ratios are extremely unbalanced or not scalable for large data analyses. To address these problems, we propose SKAT/SKAT-O type region-based tests, where the single-variant score statistic is calibrated based on SPA and Efficient Resampling (ER). Through simulation studies, we show that the proposed method provides well-calibrated p-values. In contrast, when the case-control ratio is 1:99, the unadjusted approach has greatly inflated type I error rates (90 times of exome-wide α =2.5×10^{-6}). Additionally, the proposed method has similar computation time as the unadjusted approaches and is scalable for large sample data. In our application, the UK Biobank whole exome sequence data analysis of 45,596 unrelated European samples and 791 PheCode phenotypes identified 10 rare variant associations

with p-value < 10^{-7}, including the associations between *JAK2* and myeloproliferative disease, *HOXB13* and cancer of prostate, and *F11* and congenital coagulation defects. All analysis summary results are publicly available through a web-based visual server, which can help facilitate the identification of the genetic basis of complex diseases.

2.1 Introduction

With the decreased cost of sequencing, big biobanks have started to whole exome or whole genome sequence large number of participants to identify the role of rare variants in complex diseases(Clare Bycroft et al., 2018; Dewey et al., 2016; Van Hout et al., 2019). By combining rich phenotypic information in electronic health record (EHR)(Bush et al., 2016), these sequence data will illuminate the phenome-wide association patterns of rare variants. Since most diseases and symptoms have low prevalence, the binary phenotypes in biobanks generally have unbalanced case-control ratios (1:10 or 1:100, for example)(Rounak Dey et al., 2017). For example, in the UK Biobank data, nearly 99% of PheCode-based binary phenotypes have case-control ratios less than 1:10 (Zengini et al., 2018). Substantial challenges are posed when analyzing the associations between rare variants and unbalanced phenotypes.

Since single-variant tests are underpowered to identify disease associated rare variants(Seunggeung Lee et al., 2014), gene- or region-based multiple variant tests, including the burden test(Li & Leal, 2008; Morgenthaler & Thilly, 2007), SKAT(Michael C Wu et al., 2011), and SKAT-O(Seunggeun Lee et al., 2012), are commonly used to identify rare variant associations. To evaluate the association signals in multiple variants, these methods aggregate single variant score statistics. However, as shown in our simulation studies and elsewhere(Ma et al., 2013; Wang et al., 2016; Zhang et al., 2019), these methods suffer from the inflation of type I error rates when case-control ratios are unbalanced. For single variant tests, the recently developed saddlepoint approximation (SPA) based approach provides accurate p-values under such a case-control imbalance(Rounak Dey et al., 2017; Wei Zhou et al., 2018). Although a few methods exist that adjust for unbalanced case-control ratios for gene- or region-based tests, including moment-based adjustment (MA)(Lee et al., 2015) and efficient resampling (ER)(Lee et al., 2015), these methods

are not scalable or accurate for biobank data. When the case-control ratio is extremely unbalanced, MA can still have inflated type I error rates. ER is computationally expensive when minor allele counts (MAC) are moderate or large.

To address these problems, we propose a robust region-based test that adjusts single variant score statistics using SPA and ER and aggregate the adjusted statistics. The SPA and ER help to precisely calculate the reference distribution of the single variant score statistics, thereby properly controlling for type I error rates. The computation cost of the proposed approach is comparable to unadjusted tests, and can hence be applied to large biobank data. Using extensive simulation studies, we demonstrate that our robust burden, SKAT, and SKAT-O tests have proper type I error rates even when the case-control ratio is 1:99 and exhibit larger power compared to the unadjusted burden, SKAT, and SKAT-O test. In addition, the method can be applicable not only to rare variant tests but also to the joint association test of common and rare variants.

The UK Biobank resource(Clare Bycroft et al., 2018) completed the first tranche of whole exome sequencing (WES) data for 49,960 participants(Van Hout et al., 2019). We performed robust gene-based rare-variant tests of 45,596 unrelated European samples on 791 phenotypes with at least 50 cases and identified 10 rare variant associations with p-value $< 10^{-7}$, including the associations between *JAK2* (MIM: 147796) and myeloproliferative disease (MIM: 254700), *HOXB13* (MIM: 604607) and cancer of prostate (MIM: 610997), and *F11* (MIM: 264900) and congenital coagulation defects (MIM: 134520). These results anticipate the discoveries we can make with the full 500,000 WES samples, which will be available in near future. In addition, the analysis results can be used as a community resource and facilitate the identification of the genetic basis of complex diseases.

2.2 Methods

2.2.1 Gene/region-based rare variant tests for binary traits

Assume n individuals are sequenced in a region, which has m rare variants. For the i-th individual, let y_i denote a binary phenotype, $G_i = (g_{i1}, g_{i2}, \ldots, g_{im})'$ the hard call genotypes ($g_{ij} = 0,1,2$) or dosage values of the m genetic variants in the target gene or region, and $X_i = (X_{i1}, X_{i2}, \ldots, X_{is})'$ the covariates, including the intercept. To model a binary outcome, the following logistic regression model can be used:

$$logit(\pi_i) = X_i'\alpha + G_i'\beta,$$

where π_i is the disease probability for the i-th individual, α is an $s \times 1$ vector of regression coefficients of covariates, and β is an $m \times 1$ vector of regression coefficients of genetic variants. Suppose $S_j = \sum_{i=1}^{n} g_{ij}(y_i - \hat{\pi}_i)$ is the score statistic for the variant j, where $\hat{\pi}_i$ is the estimated disease probability under the null hypothesis of no association (i.e. $\beta = 0$). Burden and SKAT test statistics can be written as

$$Q_B = \left(\sum_{j=1}^{m} \omega_j S_j\right)^2, \qquad Q_s = \sum_{j=1}^{m} {\omega_j}^2 {S_j}^2,$$

where w_j is the weight for each variant.(Michael C Wu et al., 2011) In the simulation and real data analysis, we used beta(1,25) weights, which upweight rarer variants(Michael C Wu et al., 2011). The SKAT-O method combines the burden test and SKAT with the following framework:

$$Q_\rho = (1-\rho)Q_B + \rho Q_s,$$

where ρ is a tuning parameter with range [0,1]. Since the optimal ρ is unknown, SKAT-O applies the minimum p-values over a grid of ρ as a test statistic.

Under the null hypothesis, $S = (S_1, \dots, S_m)^T$ asymptotically follows the multivariate normal distribution, $MVN\left(0,\ V^{\frac{1}{2}}CV^{\frac{1}{2}}\right)$, where C is the correlation matrix among m variants and V is a diagonal matrix where the diagonal elements are the asymptotic variances of S. In the presence of a case-control imbalance, however, the distribution of score statistics is skewed, which causes the inflation of type I error rates. To address this problem, we will utilize SPA and ER to adjust the variance matrix V.

2.2.2 Saddle Point Approximation (SPA) and Efficient Resampling (ER)

SPA is a statistical method to calculate the distribution function using the cumulant generating function (CGF). Since it utilizes all the cumulants, SPA is more accurate than using normal approximation, which only uses the first two cumulants (mean and variance). From the work of Dey et al(Rounak Dey et al., 2017), suppose $K_j(t)$ is the CGF of the score statistic S_j, which can be derived based on the fact that $Y_i \sim \text{Bernoulli}(\pi_i)$ under the null. Then, the distribution function of the score statistic S_j can be approximated by

$$\Pr(S_j < \text{s}) = \tilde{F}(s) = \Phi\left\{d + \frac{1}{d}\log\left(\frac{v}{d}\right)\right\},$$

where $d = sgn(\hat{t})\sqrt{2\left(\hat{t}s - K_j(\hat{t})\right)}$, $\text{v} = \hat{t}\sqrt{K_j''(\hat{t})}$, $\hat{t}$ is the solution to the equation $K_j'(\hat{t}) = s$, and Φ is the distribution function of the standard normal distribution(Rounak Dey et al., 2017).

Although SPA performs better than normal approximation, since it is still an asymptotic-based approach, SPA can result in inaccurate p-values when MAC is very low. To address this issue, we use ER for low MAC variants. ER is a resampling method that resamples the case–control status of individuals with a minor allele at a given variant and disease risk π_i instead of permuting case–control status across all individuals. This is because only individuals with minor

alleles contribute to the score statistics *S*. Since ER is resampling-based, it can provide an accurate p-value for a very rare variant. When MAC is low (ex. MAC ≤ 10), ER can rapidly calculate the exact p-value by numerating all possible configurations of case-control statuses. The detailed derivations of ER can be found in Lee et al(Lee et al., 2015).

2.2.3 Robust burden test, robust SKAT and robust SKAT-O

For each variant j, when the score statistic S_j lies within 2 standard deviations of the mean, the normal approximation generally performs well(Rounak Dey et al., 2017). Otherwise, due to the skewed distribution, the normal approximation causes inflated type I error rates. Hence, when S_j is beyond 2 standard deviations of the mean, we apply SPA (when MAC > 10) or ER (when MAC ≤ 10) to calculate the p-value $\tilde{p}_j$, which will be used to calibrate the variance of S_j.

Let $S_j^2/\hat{V}_j$ be a square-standardized test statistic in which $\hat{V}_j$ is the estimated variance of S_j^2. When S_j follows the normal distribution, $S_j^2/\hat{V}_j$ follows the chi-square distribution with one degree of freedom. We adjust the variance so that the p-value is the same as $\tilde{p}_j$, in which the adjusted variance is

$$\tilde{V}_j = S_j^2/\chi^2_{quantile}(1-\tilde{p}_j),$$

where $\chi^2_{quantile}$ is the quantile function of the chi-square distribution with one degree of freedom. Note that if S_j lies within 2 standard deviations of the mean, $\tilde{V}_j = \hat{V}_j$. Suppose $\tilde{V} = (\tilde{V}_1, \tilde{V}_2, \dots, \tilde{V}_m)^T$, then the p-value of the region can be calculated based on the assumption that

$$S \sim MVN\left(0,\ \tilde{V}^{\frac{1}{2}} C \tilde{V}^{\frac{1}{2}}\right).$$

These adjustments overcome the inflated type I error rates for common variants, but are insufficient to address the inflation issue for rare variants --4.87 times of exome-wide alpha=2.5×10^{-6} when the case-control ratio is 1:99. Details can be found in Table S2.1. We apply

additional adjustment by using the fact that the burden test can be presented as a single marker test with collapsed variants, and SPA performs very well for single marker test. From the above equation, the variance estimate of the burden test is $\tilde{V}_{burden} = w^T \tilde{V}^{\frac{1}{2}} C \tilde{V}^{\frac{1}{2}} w$, where $w = (w_1, \dots, w_m)^T$ is an $m \times 1$ vector of the weight. Suppose $g_i^{burden} = \sum_{j=1}^{m} w_j g_{ij}$, and then the burden test statistic (i.e. Q_B) is identical to S_{burden}^2, where $S_{burden} = \sum_{i=1}^{n} g_i^{burden}(y_i - \hat{\pi}_i)$, and the p-value $\check{p}_{S_{burden}}$ of S_{burden} can be calculated from SPA. Using the similar approximation as above, we estimate the variance S_{burden} as $\check{V}_{sum} = S_{burden}^2 / \chi^2_{quantile}(1 - \check{p}_{S_{burden}})$. Suppose $r = \tilde{V}_{sum} / \check{V}_{sum}$. In order to control type I error inflation, we suggest utilizing a more conservative variance. Let $\tilde{r} = \min(1, r)$, then

$$S \sim MVN\left(0,\ \left(\frac{\tilde{V}}{\tilde{r}}\right)^{\frac{1}{2}} C \left(\frac{\tilde{V}}{\tilde{r}}\right)^{\frac{1}{2}}\right).$$

With this formula, Robust burden, SKAT and SKAT-O tests can be performed.

2.2.4 Extension to the joint test of common and rare variants

Our robust method can be extended to the joint test of common and rare variants. Consider the following model

$$logit(\pi_i) = X_i'\alpha + G_{1i}'\beta_1 + G_{2i}'\beta_2.$$

For the individual i, π_i is the disease probability; X_i is the vector containing all the covariates, including the intercept; G_{1i} is the genotype vector of rare variants with length m_r; and G_{2i} is the vector of common variants with length m_c. To test the hypothesis of no genetic effects $H_0: \beta_1 = 0,\ \beta_2 = 0$, the test statistic Q_ϕ can be written as

$$\begin{aligned} Q_\phi &= (1-\phi)Q_{rare} + \phi Q_{common} \\ &= (1-\phi)S_1' W_1 W_1' S_1 + \phi S_2' W_2 W_2' S_2, \end{aligned}$$

where S_1 and S_2 are the vectors of score statistics for rare and common variants respectively, and W_1 and W_2 are diagonal weight matrices for rare and common variants.

Under the null, $S = (S_1,\ S_2) \sim MVN\left(0,\ V^{\frac{1}{2}} C V^{\frac{1}{2}}\right)$. Using the approach described in the previous section, we apply SPA and ER to calibrate variance estimates to perform a robust SKAT method.

2.2.5 Numerical Simulations

We conducted extensive simulation studies to evaluate the performance of the proposed methods for dichotomized traits. The sequence data of mimicking European ancestry over 200 kb regions were generated using the calibrated coalescent model(Schaffner et al., 2005). We randomly selected regions with lengths of 1, 2, and 3 kb and tested for associations in all simulation settings. On average each simulated dataset had 16.33 (SD: 4.05), 32.69 (SD: 5.65) and 49.05 (SD: 6.71) rare variants for 1, 2, and 3 kb regions, respectively, when the sample size was 50,000.

We generated data sets with sample size 50,000. We included two covariates for the analysis. The first one followed a Bernoulli distribution with $p = 0.5$ and the other followed the standard normal distribution, corresponding to the gender and normalized age. Four case-control ratios were considered, 1:1, 1:9, 1:49 and 1:99, and the binary phenotypes were simulated from

$$\text{logit}(\pi_i) = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \beta_1 g_{1i} + \cdots + \beta_m g_{mi},$$

where $\beta_1 = \beta_2 = \cdots = \beta_m = 0$; γ_1 and γ_2 were chosen to let the odds ratio (OR) of X_1 and X_2 equal 1.2 and 1.5 respectively, and γ_0 was chosen based on disease prevalence. Seven different methods were applied to each of the generated datasets. For all variants in the region, we applied the unadjusted and robust joint test of common and rare variants. For rare variant tests (MAF<=0.01), we applied (1) burden test; (2) robust burden test; (3) SKAT; (4) robust SKAT; (5) SKAT-O; (6) robust SKAT-O; and (7) the hybrid method. The hybrid method(Lee et al.,

2015), developed by Lee, selects a method among ER, Quantile adjusted moment matching (QA) and Moment matching adjustment (MA) based on MAC, and the degree of case-control imbalance. A total of 10^7 phenotypes were generated, and type I error rates were estimated by the proportion of p-values smaller than the given α level divided by given α.

For power simulations, 30% of variants were randomly selected as causal. Two settings were considered: (1) 80% causal variants were risk-increasing variants and 20% were risk-decreasing variants; and (2) all causal variants were risk-increasing variants. For each setting, 10,000 data sets were generated, and the power was estimated as the proportion of p-values smaller than the empirical α level, which was calculated in the type I error simulation.

2.2.6 Analysis of whole exome sequencing (WES) data in the UK Biobank

We analyzed the first tranche of UK Biobank WES data with 49,960 participants(Van Hout et al., 2019). Due to the quality issues in the Regeneron pipeline(Biobank, 2019), we analyzed genotype data processed from the FE pipeline(Regier et al., 2018). The details of sample selection and QC procedures are described elsewhere[1]. We excluded one individual in related pairs (up to second-degree relatives) to identify a set of unrelated individuals. To preserve cases, we first selected a maximal set of unrelated cases, then removed controls that were related to the unrelated cases and kept a maximal set of unrelated controls. Because of the missing values in the phenotypes, the individuals included in the analysis varied across phenotypes. We performed gene-based tests on 45,596 independent European participants in the UK Biobank, whose phenotype data were available.

With a previously published scheme(Denny et al., 2013), we defined disease-specific binary phenotypes by combining hospital ICD-9 codes into hierarchical PheCodes, each representing a specific disease group. ICD-10 codes were mapped to PheCodes using a combination of

available maps through the Unified Medical Language System, manual review and other sources. Study participants were labeled a PheCode if they had one or more of the PheCode-specific ICD codes. Cases were defined as all study participants with the PheCode of interest and controls were all study participants without the PheCode of interest. Gender checks were performed, so PheCodes specific for one gender could not be assigned to the other gender by mistake(Wei Zhou et al., 2018).

There were 791 binary phenotypes with at least 50 cases based on PheCodes, in which 551 phenotypes had case-control ratios smaller than 1:99. Because our robust methods would cause a certain inflation for extremely unbalanced case-control ratios (Table S2.1), and using more controls than those from case-control ratio of 1:99 would not improve power (Figure S2.1), we did matching on these 551 traits using the first 4 genotype principal components. Specifically, for each case we found the closest controls in Euclidean distance to make the case-control ratio be 1:99. We used principal components calculated by UK-Biobank, which were calculated from 147,551 LD-pruned SNPs with missing rate < 0.015 and MAF > 0.01(Bycroft et al., 2017).

We focused on the rare variants (MAF<=0.01) of the nonsynonymous and splicing variants in the exon and neighboring regions. Particularly we used annotation of frameshift deletion, frameshift insertion, nonframeshift deletion, nonframeshift insertion, nonsynonymous SNV, splicing, stopgain and stoploss from ANNOVAR (Version built on 2018-04-16) with refGene database (hg38)(Wang et al., 2010). A total of 18,360 genes were used for the analysis. The number of variants in genes ranged from 2 to 7,439 with a highly skewed distribution (Figure S2.2). The six methods discussed in the simulation study, unadjusted and robust versions of the burden test, SKAT and SKAT-O methods, were applied to the data. Age, gender and the first four principal components were used as covariates to adjust for population stratification.

2.3 Results

2.3.1 Type I Error and Power Simulation Results

We generated 10^7 datasets to compare type I error rates of the proposed approaches (robust burden, SKAT and SKAT-O), unadjusted approaches (burden, SKAT and SKAT-O) and a hybrid approach for SKAT-O(Lee et al., 2015). The hybrid approach applies several adjustment methods based on MAC. Table 2.1 shows that the unadjusted approaches had substantial inflation of type I error rates when the case-control ratio was unbalanced and the region length was 1 kb. In contrast, the robust approaches controlled type I error rates much better and had only a slight inflation when the case-control ratio was 1:99. Interestingly, the existing hybrid approach showed substantially inflated type I error rates when case-control ratios were extremely unbalanced (case-control ratio=1:49 and 1:99). This may be due to the fact that the MAC-based method selection rule in the hybrid approach does not perform well under extremely unbalanced case-control ratios. When the case-control ratios are more extreme than 1:99, the robust SKAT and SKAT-O showed some inflation of type I error rates (Table S2.1). Simulation studies with 2 and 3 kb regions show that empirical type I error rates of robust SKAT-O are generally similar regardless of region length (Table S2.2). Additionally, when testing both common and rare variants, robust SKAT can control type I error rates well compared with unadjusted SKAT (Table S2.3). Overall, the type I error simulation results confirmed that the proposed robust approaches provide substantially improved type I error rates compared to the unadjusted and existing hybrid approaches.

Table 2.1. Type I error rates of unadjusted and robust versions of burden, SKAT and SKAT-O and hybrid method. Total 10^7 datasets of 1 kb regions were generated to estimate type I error rates. Each cell represents an empirical type I error rate divided by significance level α. The sample size was 50,000.

α	Case: Control	Burden	Robust burden	SKAT	Robust SKAT	SKAT-O	Robust SKAT-O	Hybrid SKAT-O
10^{-2}	1:1	1.00	1.00	0.99	0.99	1.11	1.11	1.09
	1:9	0.99	1.00	1.01	1.01	1.13	1.13	1.09
	1:49	1.02	0.95	1.44	1.22	1.44	1.23	1.27
	1:99	1.07	0.91	1.92	1.41	1.82	1.33	1.53
10^{-4}	1:1	1.02	1.00	0.99	1.03	1.27	1.32	1.27
	1:9	1.12	0.99	1.39	1.14	1.65	1.40	1.52
	1:49	2.43	0.97	6.31	1.65	6.16	1.79	4.54
	1:99	3.95	1.02	13.48	2.13	12.77	2.17	8.89
	1:1	1.11	1.03	1.24	1.54	1.38	1.38	1.40
2.5	1:9	1.29	0.77	2.47	1.45	2.51	1.49	2.23
$\times 10^{-6}$	1:49	6.88	1.06	28.27	1.91	23.70	1.98	16.69
	1:99	16.34	0.90	89.53	1.81	71.32	1.60	42.59

Figure 2.1 shows the empirical powers of the hybrid, unadjusted and robust versions of SKAT-O methods, when 80% causal variants were risk-increasing variants and 20% were risk-decreasing variants. The empirical powers of unadjusted and robust versions of the burden tests and SKAT can be found in Figure S2.3. Since unadjusted and hybrid methods had severely inflated type I error rates, for the fair comparison, we used the empirical significance level estimated from type I error simulation studies. Assuming that the type I error rates could be properly controlled for all methods, robust SKAT-O had similar power as unadjusted SKAT-O in balanced and moderately unbalanced case-control ratios (1:1 and 1:9) and was more powerful than unadjusted SKAT-O in extremely unbalanced ratios (1:49 and 1:99). Robust burden tests had the same power as unadjusted burden tests across all four case-control ratios. Robust SKAT had similar power as unadjusted SKAT in balanced ratios and was more powerful than unadjusted SKAT in unbalanced ratios. If the number of cases was fixed, more controls (1:49 and 1:99) increased power greatly compared to case-control ratio 1:1 for all three robust methods (Figure S2.1). In addition, we found that 1:99 had slightly more power than 1:49, where we could infer that 1:99 is sufficient to achieve the maximum power and more controls can hardly increase the power. The

power simulation results with different region lengths (Figure S2.4) and power simulation results with all causal variants being risk-increasing variants (Figure S2.5) were quantitatively similar. In summary, the robust methods had similar or more power than the unadjusted methods in all scenarios. Among the three robust methods, robust SKAT-O generally performed better than robust SKAT and robust burden tests since robust SKAT-O combined the two tests (Figure S2.6).

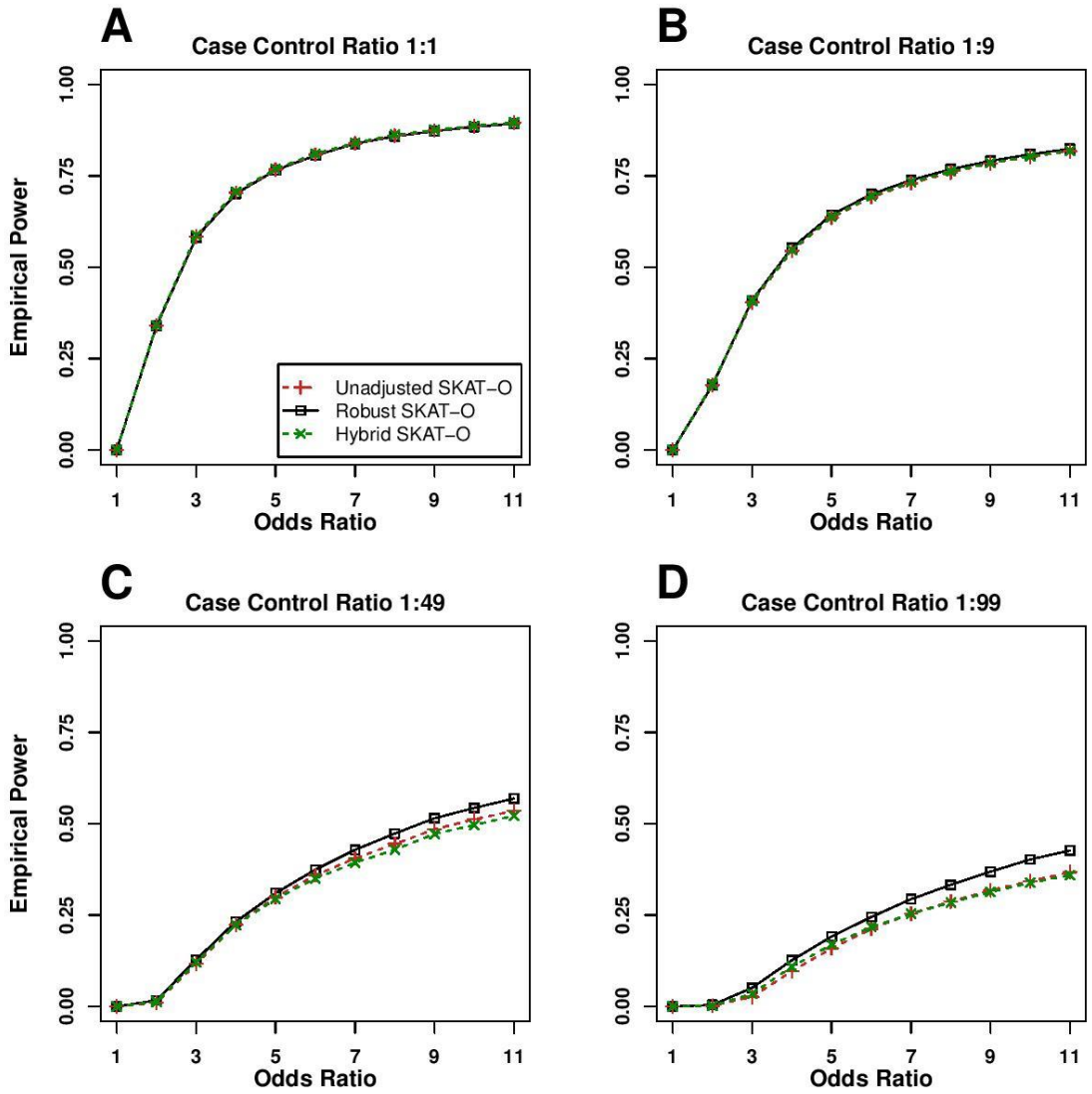


Figure 2.1 Empirical power estimates for the unadjusted and robust versions of SKAT-O, and hybrid method. Power was calculated at the empirical α levels estimated from Type I error simulations with adjusting type I error rate at 2.5×10^{-6}. Total 10,000 datasets were generated

with region length 1kb. 30% of variants were causal variants with 80% causal variants being risk-increasing and 20% being risk-decreasing. The sample size was 50,000. The X-axis represents the genetic effect odds ratio and the Y-axis represents the empirical power.

2.3.2 Comparison of computational times

To compare the computation times, we generated 1,000 datasets (Figure 2.2). Since SKAT-O combines the burden and SKAT tests, we only considered the SKAT-O test. As the sample sizes increased, the computation time of ER increased and required ~16.1 CPU hours for analyzing one gene for 50,000 individuals. In contrast, unadjusted methods required 140x less computation time (~6.7 min) and the computation times barely changed by sample size (5,000-100,000 individuals). Our robust method performed similarly as unadjusted SKAT-O (~8.5 min). Since the hybrid approach selects its methods based on MAC and case-control ratios, the computation cost of the hybrid approach is not determined by the sample size. Overall, the hybrid approach was slower than the proposed method. The computation time for analyzing UK-Biobank data of 791 binary phenotypes with robust SKAT-O was 453 CPU days, i.e. ~13.7 CPU hours per one phenotype.

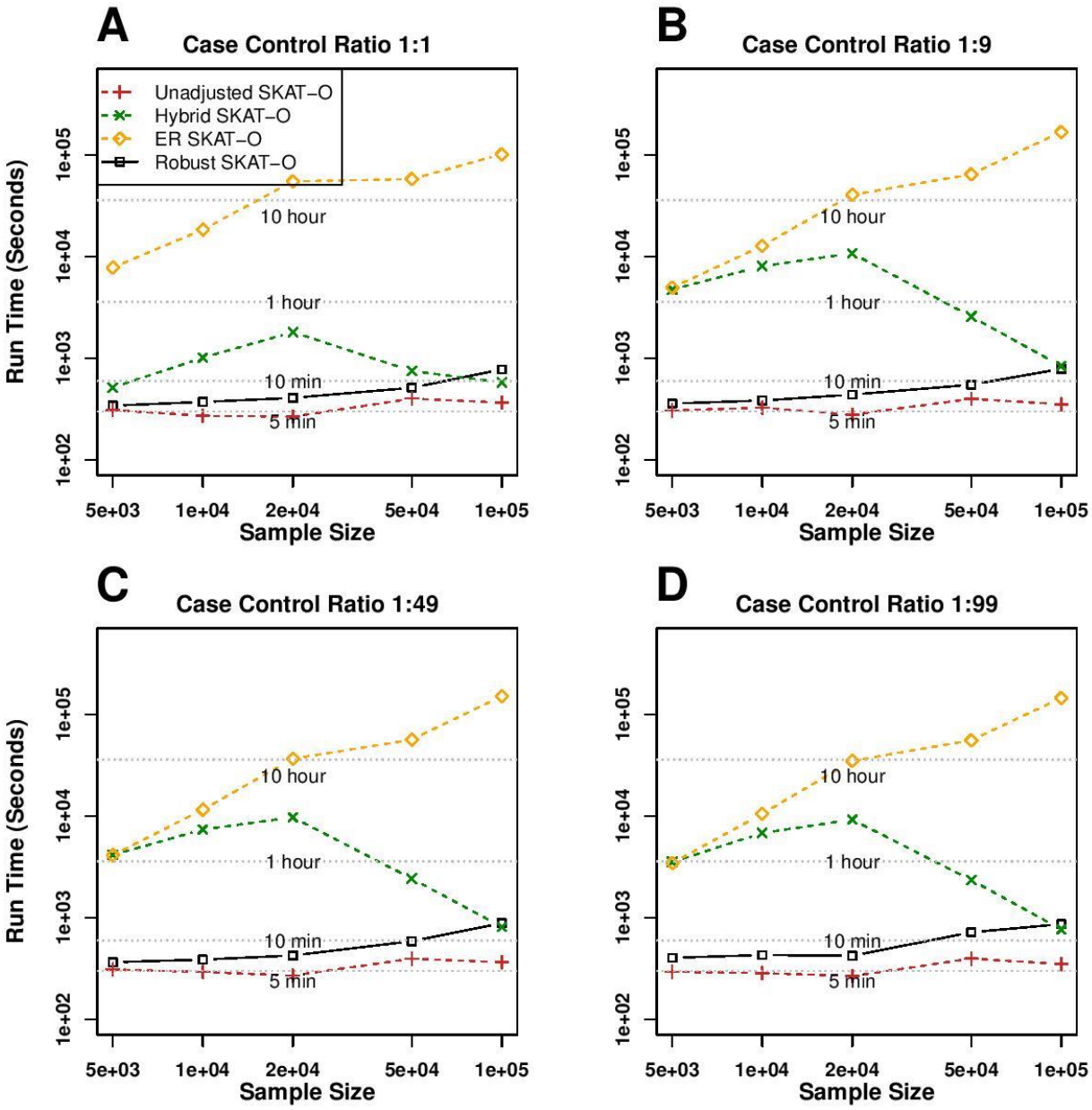


Figure 2.2 Comparison of computation time of unadjusted, hybrid, ER and robust approaches for SKAT-O.
The rare-variant region-based tests were performed on randomly selected 1 kb regions of 1,000 resamples. The X-axis represents the sample size and the Y-axis represents the run time of 1,000 resamples.

2.3.3 Analysis of whole exome sequencing (WES) data in the UK Biobank

We applied six methods (unadjusted and robust versions of burden, SKAT and SKAT-O) to the analysis of WES data in the UK Biobank. We restricted our analysis to the rare nonsynonymous

and splicing variants with minor allele frequencies (MAFs) < 0.01 in exon regions. A total of 18,360 genes were analyzed based on 45,596 independent European samples across 791 binary phenotypes with at least 50 cases. For phenotypes with case-control ratios more extreme than 1:99, we identified the ancestry-matched control samples to make the case-control ratios 1:99 (See Methods).

With the cutoff of $\alpha = 2.5 \times 10^{-6}$, unadjusted SKAT-O detected 73,723 significant associations, most of which would be false positives, while our robust methods detected 34 significant associations for the burden test, 99 for SKAT and 111 for SKAT-O (Table S2.4). Since we were testing many phenotypes, the usual exome-based cutoff of 2.5×10^{-6} can produce spurious associations. Following Hout et al(Van Hout et al., 2019), we used a more stringent level $\alpha = 10^{-7}$ and identified that 10 gene-phenotype pairs had robust SKAT-O p-values smaller than 10^{-7} (Table 2.2). Among 10 phenotype-gene pairs, only 2 had a single SNP p-value $< 5 \times 10^{-8}$, indicating that gene/region-based approaches are more powerful than single variant analyses. For each gene, the top 3 smallest p-value variants were reported in Table S2.5 and single variant p-values are presented in Figure S2.7. QQ plots for those 10 phenotypes show that unadjusted SKAT-O had greatly inflated type I error rates, but our robust approach provided relatively well-calibrated results (Figure S2.8).

Rare variant associations between *JAK2* and myeloproliferative disease (number of cases=94)(Baxter et al., 2005), and *HOXB13* (MIM: 604607) and prostate cancers (MIM: 610997) (number of cases=741)(Ewing et al., 2012) have been previously reported, which demonstrates that our analysis can replicate known signals, even when the number of case samples is very small. PheWAS plot of *HOXB13* shows that there is an additional association signals between *HOXB13* and Carditis (p-value=9.18×10^{-6}) (Figure 2.3), and this may be due to

the fact that Carditis is a complication of prostate cancer biopsy and treatment(Aubry et al., 2013).

Table 2.2 Significant gene-phenotype associations in the UK Biobank WES data. Lowest P SNP means the lowest p-value of all single variants contained in the gene-phenotype association. Conditional P-value (SKAT-O) means the robust SKAT-O p-value after conditioning on the most significant nearby common variant (± 100 Kbp up and down stream). P-value of the most significant nearby variant was from SAIGE single variant analysis results(Wei Zhou et al., 2018) of the UK-Biobank imputed datasets of 400,000 British samples.

Phenotype (PheCode)	Gene Name	Case: control	The number of snps	Case MAC	Control MAC	Robust SKAT-O P-values	Lowest P SNP	Conditional P-value (SKAT-O)	P-value of the most significant nearby variant
Myeloproliferative disease (200)	*JAK2*	94:9306	73	27	442	1.36E-33	1.81E-41	1.06E-35	2.30E-17
Unspecified monoarthritis (716.2)	*OGG1*	1728:41060	117	118	1643	7.73E-09	4.67E-04	7.79E-09	4.28E-04
Menopausal and postmenopausal disorders (627)	*NFE2L3*	1345:21226	171	145	1358	2.54E-08	2.72E-05	3.94E-08	2.14E-04
Cancer of prostate (185)	*HOXB13*	741:18940	37	18	154	3.00E-08	5.24E-08	2.50E-08	1.17E-04
Other aneurysm (442)	*P3H1*	164:16236	110	17	497	5.76E-08	1.71E-05	4.03E-07	1.22E-03
Heartburn (530.9)	*USP45*	189:18711	103	24	649	6.34E-08	5.39E-05	1.46E-09	4.08E-02
Fracture of hand or wrist (804)	*GSDMC*	382:37818	109	25	761	7.12E-08	8.17E-05	1.49E-07	1.84E-02
Congenital coagulation defects (286.1)	*F11*	76:7524	38	8	84	7.40E-08	4.52E-05	4.09E-08	6.30E-03
Congenital anomalies of great vessels (747.13)	*SLC46A1*	134:13266	28	11	255	9.38E-08	1.86E-08	3.87E-08	2.29E-03
Peptic ulcer (excl. esophageal) (531)	*LMNB2*	773:44818	171	24	508	9.89E-08	3.83E-06	9.54E-08	1.31E-03

Among other genes, *P3H1* (MIM: 610339), Prolyl 3-Hydroxylase 1, was observed to be associated with other aneurysm (p-value=5.76×10^{-8}) and possibly associated with abdominal aortic aneurysm (p-value=5.79×10^{-5}). *P3H1* is involved in collagen metabolism and was found to be present in the pulmonary artery(Vranka et al., 2004). *F11* (MIM: 264900), also known as Coagulation Factor XI, was observed as associated with congenital coagulation defects (p-value=6.13×10^{-8}), which is consistent with the fact that Factor XI participates in blood coagulation as a catalyst in the conversion of factor IX to factor IXa in the presence of calcium

ions(Asakai et al., 1987). *SLC46A1* (MIM: 611672), which encodes a transmembrane folate transporter protein, is associated to congenital anomalies of great vessels (p-value=9.38×10^{-8}), consistent with the role of folate in cardiovascular disease(Verhaar et al., 2002). PheWAS shows the close association between *SLC46A1* and two other blood diseases: cardiac congenital anomalies (p-value=9.16×10^{-7}) and cardiac and circulatory congenital anomalies (p-value=1.44×10^{-6}).

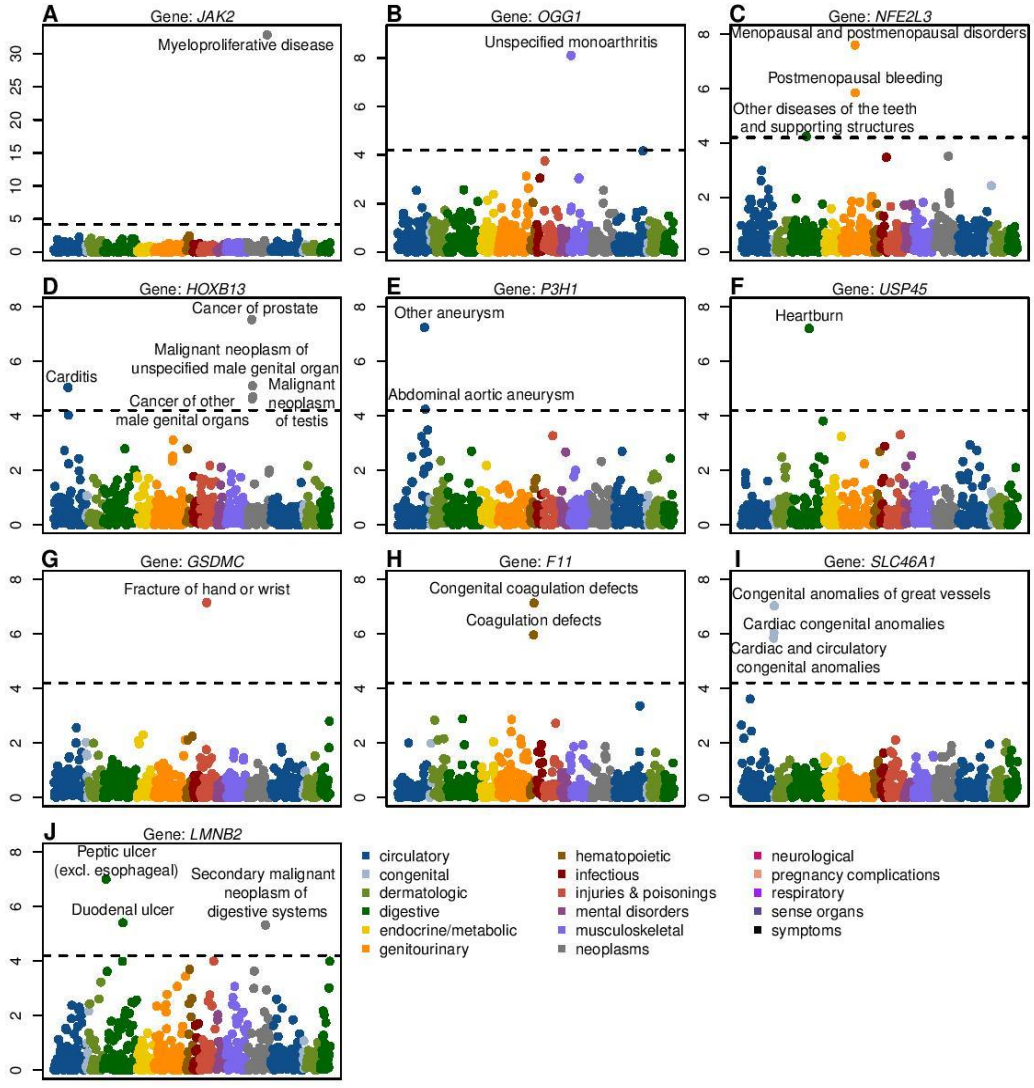


Figure 2.3 PheWAS plots of 10 rare variant associations with p-value< 10^{-7}.
The X-axis represents 791 binary traits and the Y-axis represents the negative log10 p-values. The dashed line represents the cutoff of $0.05/791=6.32\times10^{-5}$.

We carried out conditional analysis to evaluate whether the rare variant association signals were independent of the nearby common variant association signals (± 100 Kbp up and down stream) (Table S2.6). To identify most significant nearby variants, we used SAIGE single variant analysis results of the UK-Biobank imputed datasets of 400,000 British samples(Wei Zhou et al., 2018). All ten associations remained significant after the conditional analysis (Table 2.1).

We have generated summary statistics for all gene-phenotype association results using our robust approach and made them available in a PheWEB-like visual server (See Web Resources).

2.4 Discussion

In this paper, we present a robust approach that can address case-control imbalance in region-based rare variant tests. The proposed approach uses recently developed ER and SPA to calibrate the variance of single variant score statistics to accurately calculate region-based p-values. Computation cost of the proposed approach is similar to the unadjusted approach, which makes it scalable for large analysis. Simulation studies show that unadjusted methods suffer severe inflation of type I error rate in unbalanced case-control ratios while robust methods can successfully address it. The UK-Biobank exome data analysis shows that the method provides calibrated p-values and contribute to identifying true association signals.

The proposed robust methods combine SPA and ER to recalibrate variances of single score statistics. SPA can be thought as higher order asymptotic approach with error bound $O(n^{-3/2})$ (Rounak Dey et al., 2017), where *n* is the sample size, which is much smaller than the error bound of normal approximation, $O(n^{-1/2})$. However, SPA is still asymptotic-based and cannot perform well when MAC is small. Since ER is a resampling-based approach and can calculate the exact p-value when MAC is small, it can complement SPA.

Our UK Biobank WES data analysis of 45,596 European samples and 791 binary phenotypes have identified 10 rare variant associations with p-value < 10^{-7}, including the replication of two known signals. Currently UK-Biobank is carrying out whole exome sequencing for 500,000 individuals. Our analysis presents an early snapshot of the discoveries that can be made with full UK-Biobank samples.

All the UK-Biobank analysis summary statistics are publicly available, which can be a useful community resource to show detailed results of the UK-Biobank. Due to the large scale of the data, for labs not specialized in big data analysis, it is very challenging to analyze UK-Biobank exome data. The analysis results will make the data more accessible and facilitate the identification of the genetic basis of complex diseases. For example, researchers could utilize our results for meta-analysis to combine samples with different studies. They can also be used to validate novel signals from other studies.

There are several limitations in the proposed method. Currently, the robust methods require that all individuals are unrelated. Restricting analysis to unrelated samples reduce sample size and case counts in many situations(Bi et al., 2019). For example, some rare phenotypes within a health system may be clustered in a few families. Analysis based on independent samples may significantly decrease the power. When there are related individuals, generalized linear mixed model (GLMM) based approaches(Han Chen et al., 2016; Wei Zhou et al., 2018) should be used to incorporate the relatedness. Recently Chen et al developed efficient mixed effect model approach for gene-based tests(Han Chen et al., 2019) and Zhou et al expanded scalable single-variant GLMM to gene-based tests that can handle the full size of UK-Biobank data of 500,000 samples(Zhou et al., 2019). Since these methods are also based on single-variant score statistics, the robust approach can be applied to them with modifications for GLMM. We leave it for a

separate work. Second, when the case-control ratios are more extreme than case: control=1:99, the method suffers type I error inflation. Because of this, our UK-Biobank exome analysis used the matching scheme in which if the case control ratios are more extreme than 1:99, we use the matching to reduce the number of controls. Third, novel findings are not validated from independent datasets, so we cannot rule out the possibility that they are false positives. Lack of replication can be alleviated as more sequencing studies are conducted in biobanks.

In summary, we have proposed a robust region-based method and showed that the method can accurately analyze UK-Biobank exome data. With the continuous decrease of sequencing cost and growing effort to build large biobanks and cohorts(Collins & Varmus, 2015), rare variants association analysis will be increasingly applied to binary phenome. Our method will provide accurate results for binary phenome analysis and contribute to identifying the role of rare variants in complex diseases. (Hawkins & O'Doherty, 2011)

Web Resources

OMIM, http://www.omim.org

Unified medical language system, https://www.nlm.nih.gov/ research/umls

Robust gene-based test, https://github.com/leeshawn/SKAT/tree/Sparse_Version

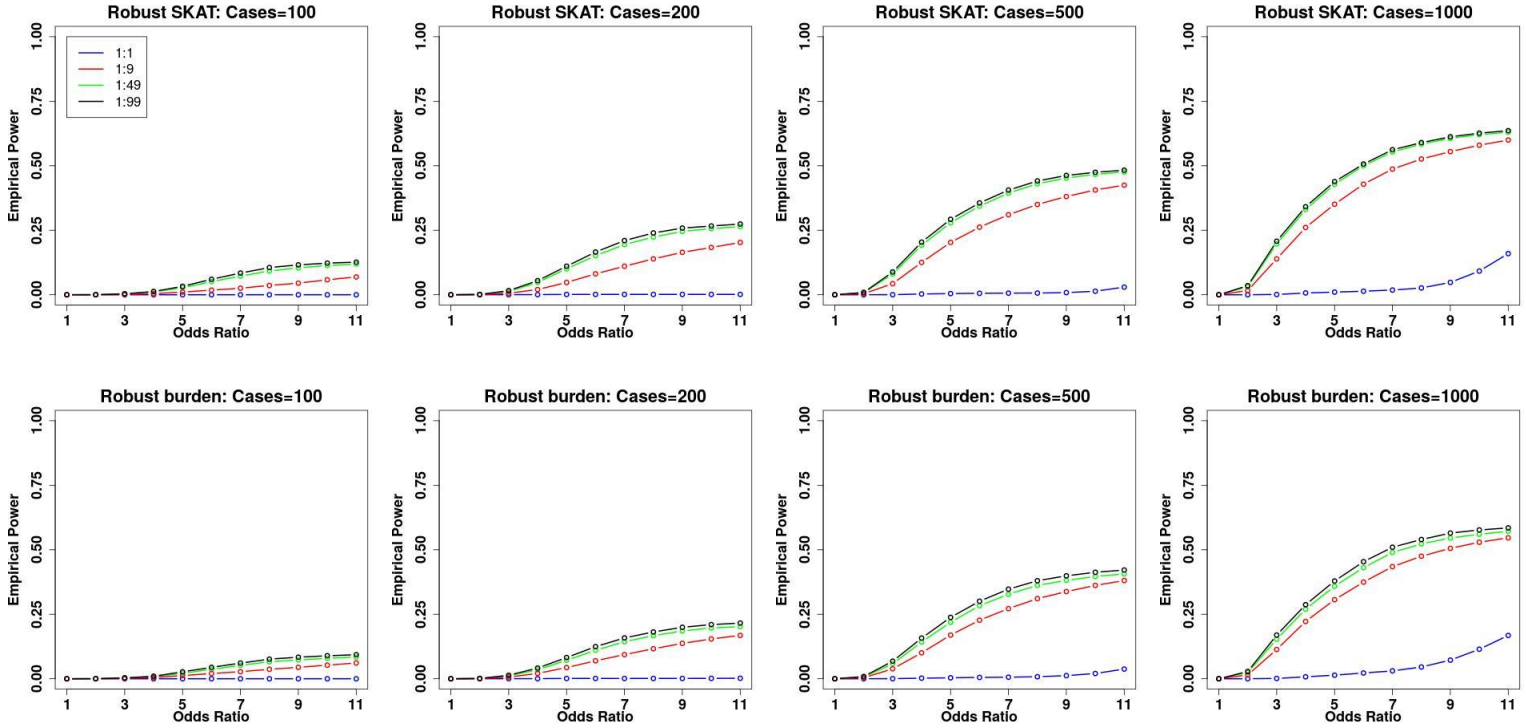
SKAT (version 1.3.2.1), https://cran.r-project.org/web/packages/SKAT

UK-Biobank, https://www.ukbiobank.ac.uk/

UK-Biobank analysis results (gene-based test for binary phenome), http://ukb-50kexome.leelabsg.org/

2.5 Supplementary Materials

2.5.1 Supplemental Figures



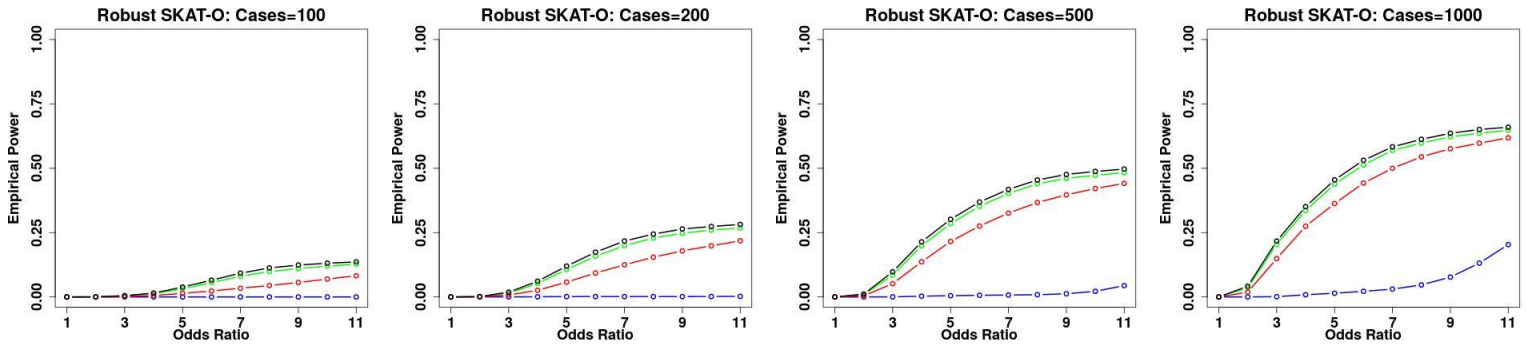


Figure S2.1. Empirical power estimates for robust SKAT, burden, SKAT-O with the same number of cases across different case control ratios. 30% of variants were causal variants and all causal variants were risk-increasing. The X-axis represents the genetic effect odds ratio and the Y-axis represents the empirical power. All causal variants had the same odds ratios.

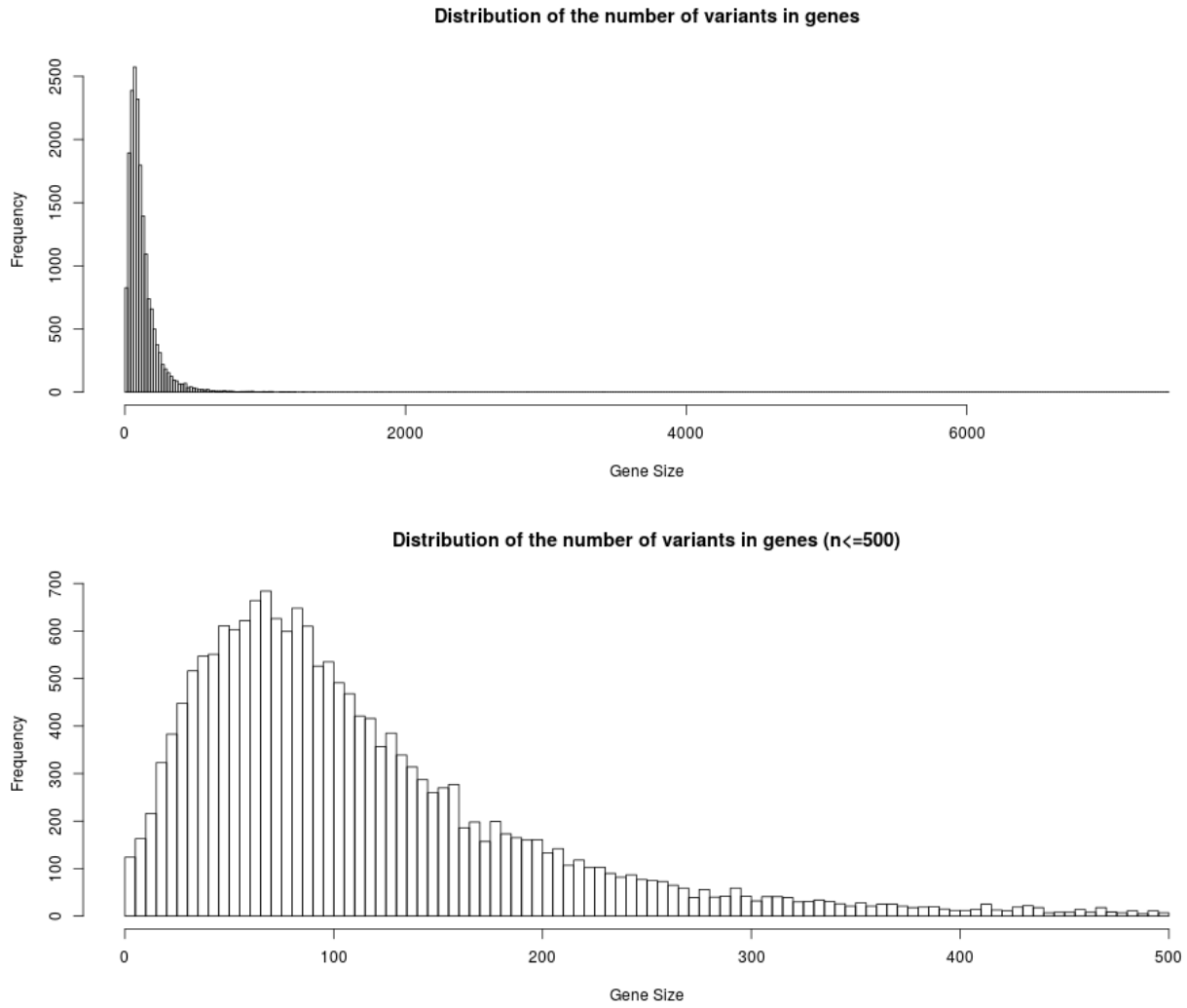


Figure S2.2. The distribution of the number of variants in genes in the UK-Biobank WES data.

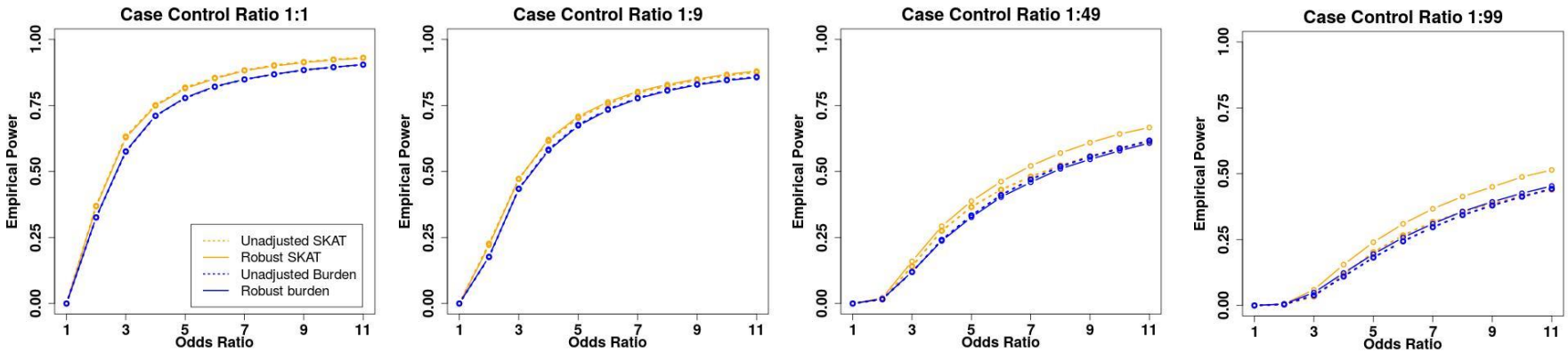
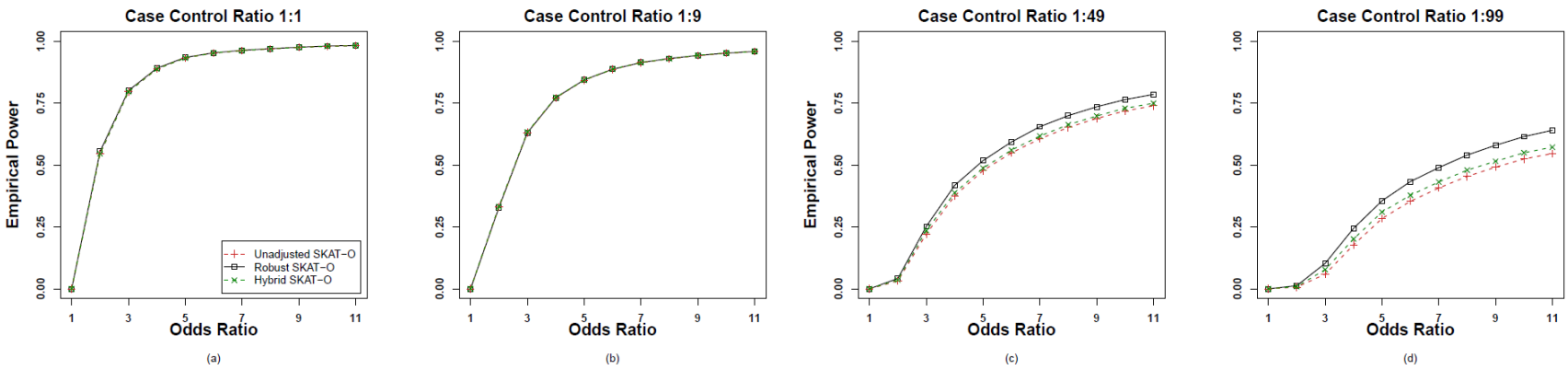


Figure S2.3. Empirical power estimates for the unadjusted and robust version of SKAT and burden test where 30% of variants were causal variants and all causal variants were risk-increasing. The X-axis represents the genetic effect odds ratio and the Y-axis represents the empirical power. All causal variants had the same odds ratios.

(A) Region length 2kb



(B) Region length 3kb

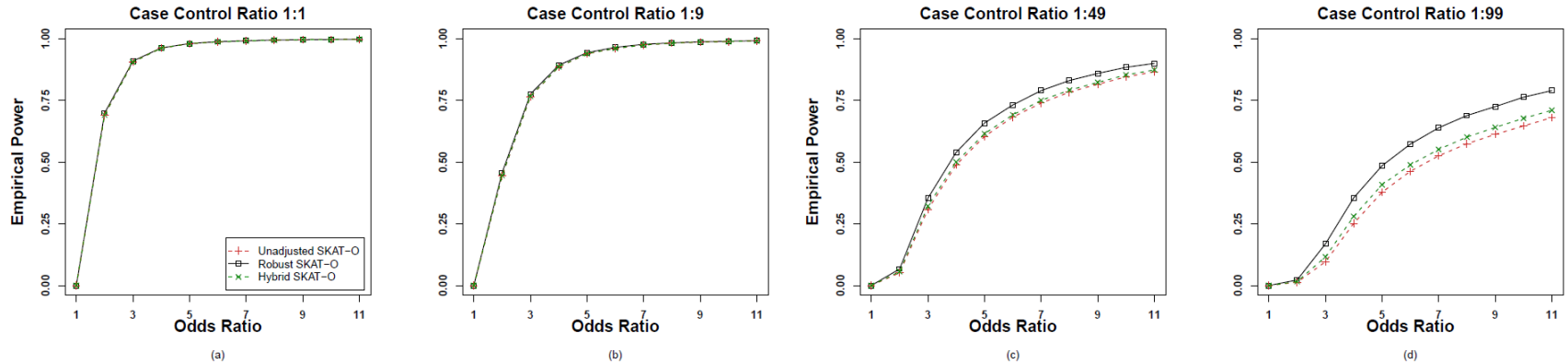


Figure S2.4. Empirical power estimates for the unadjusted and robust versions of SKAT-O and hybrid method where 30% of variants were causal variants. 80% causal variants were risk-increasing and 20% were risk-deceasing. The sample size was 50,000 and 10,000 datasets were generated. The X-axis represents the genetic effect odds ratio and the Y-axis represents the empirical power. (A) Region length is 2kb. (B) Region length is 3kb.

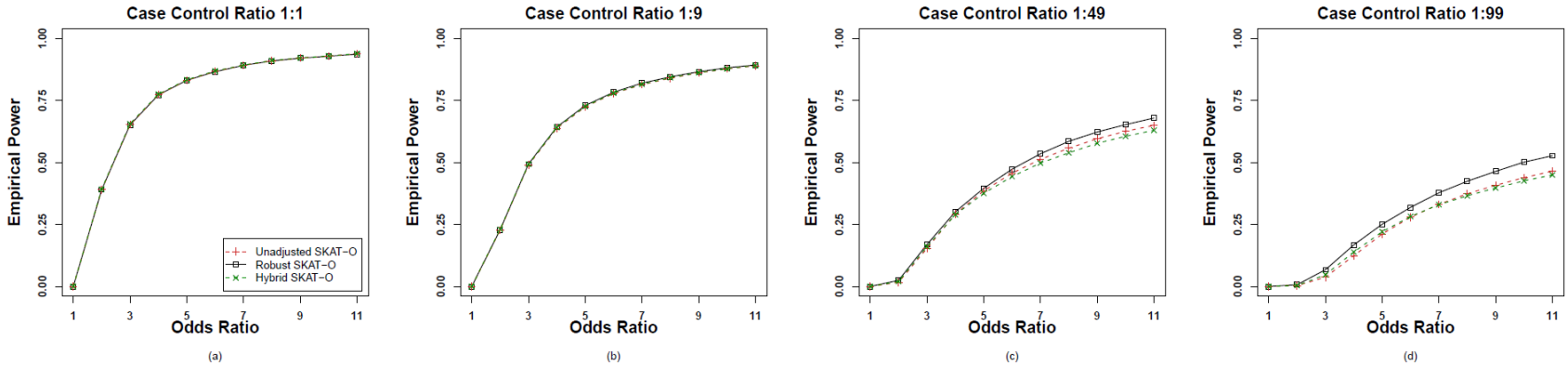


Figure S2.5. Empirical power estimates for the unadjusted and robust versions of SKAT-O and hybrid method where 30% of variants were causal variants. All causal variants were risk-increasing. The sample size was 50,000 and 10,000 datasets were generated with region length 1kb. The X-axis represents the genetic effect odds ratio and the Y-axis represents the empirical power.

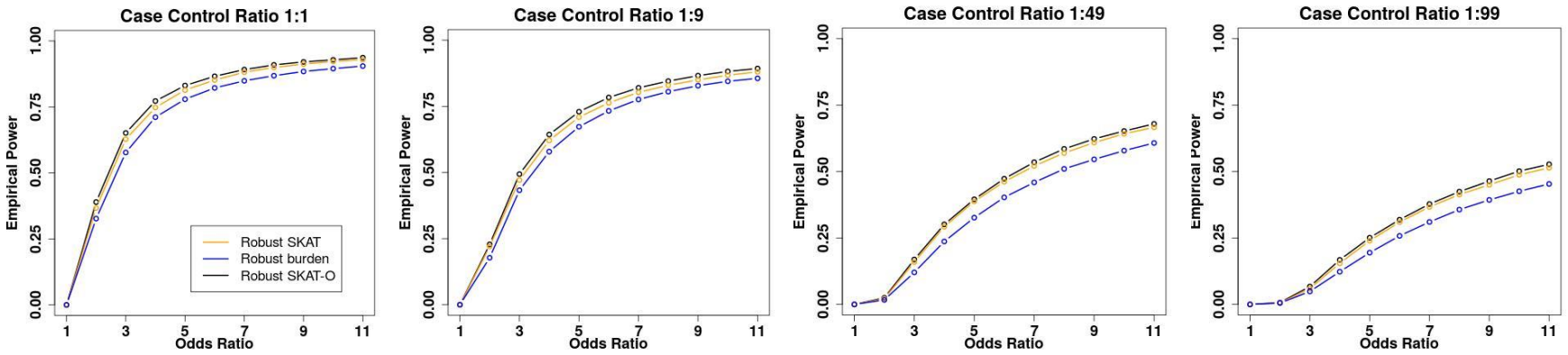


Figure S2.6. Empirical power estimates for robust SKAT, burden and SKAT-O where 30% of variants were causal variants and all causal variants were risk-increasing. The X-axis represents the genetic effect odds ratio and the Y-axis represents the empirical power. All causal variants had the same odds ratios.

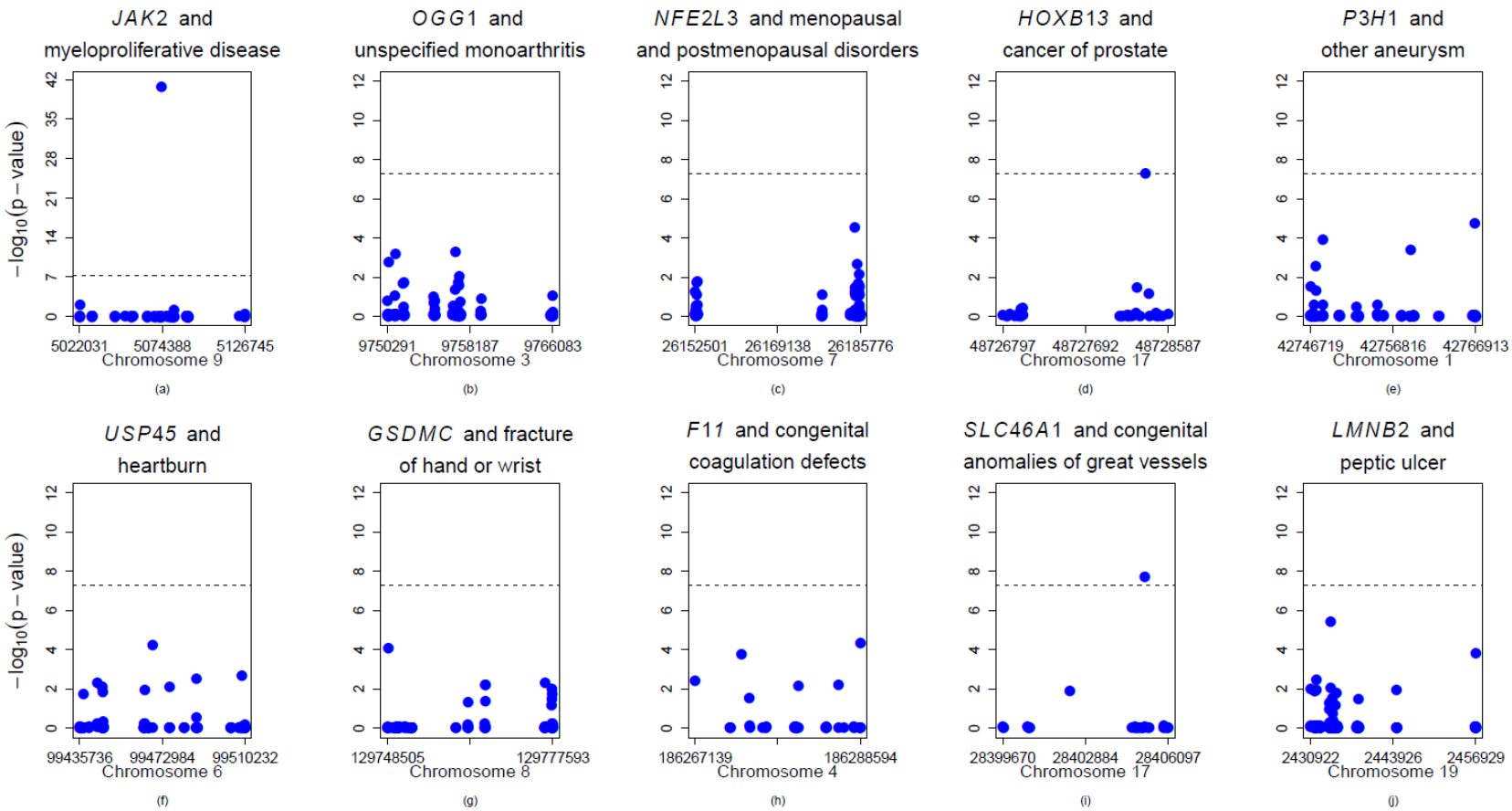


Figure S2.7. P-values of single variants in 10 significant genes. The X-axis represents the position of each single variant in the gene, and the Y-axis represents the negative log10 p-values of single variants. The dashed line represents the cutoff of 5×10^{-8}.

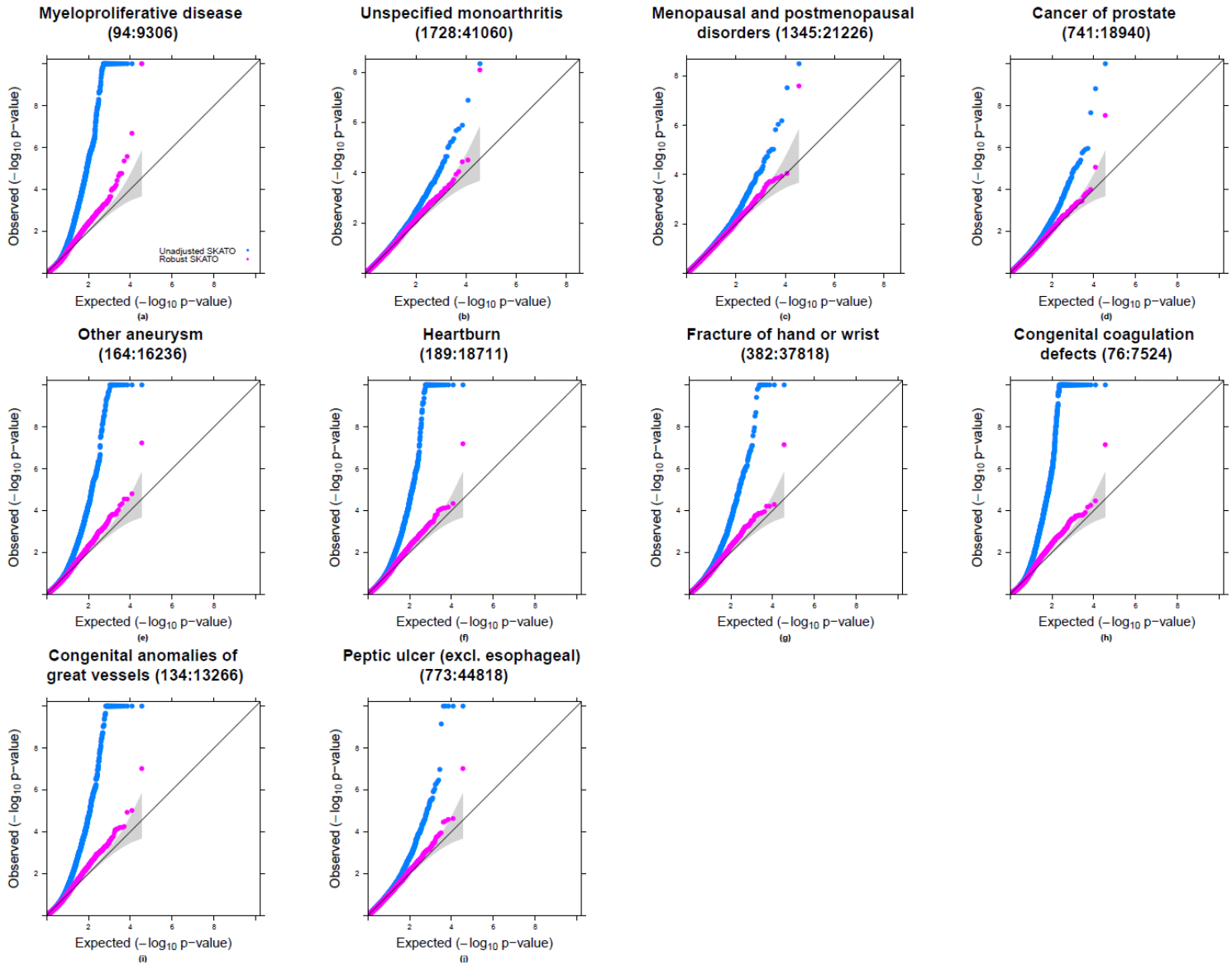


Figure S2.8. QQ Plots of SKAT-O p-values of 10 selected phenotypes. The X-axis represents the expected negative log10 p-values and the Y-axis represents the observed negative log10 p-values of genes.

2.5.2 Supplemental Tables

Table S2.1. Type I error rate divided by α of different methods when testing an association with dichotomous traits at stringent α levels $\alpha = 10^{-2}, 10^{-4}$ and 2.5×10^{-6}. The sample size was 50,000 and 10^7 datasets were generated.

α	Case: Control	SKAT	Robust SKAT without additional adjustment	Robust SKAT	Burden	Robust burden without additional adjustment	Robust burden	SKAT-O	Robust SKAT-O without additional adjustment	Robust SKAT-O	Hybrid SKAT-O
10^{-2}	1:1	0.99	0.99	0.99	1.00	1.00	1.00	1.11	1.11	1.11	1.09
	1:9	1.01	1.01	1.01	0.99	0.99	1.00	1.13	1.13	1.13	1.09
	1:49	1.44	1.24	1.22	1.02	0.95	0.95	1.44	1.24	1.23	1.27
	1:99	1.92	1.44	1.41	1.07	0.92	0.91	1.82	1.36	1.33	1.53
	1:199	2.74	1.76	1.71	1.19	0.91	0.88	2.47	1.56	1.52	2.00
	1:399	3.99	2.22	2.14	1.44	0.94	0.89	3.47	1.88	1.82	2.75
10^{-4}	1:1	0.99	0.98	1.03	1.02	1.03	1.00	1.27	1.28	1.32	1.27
	1:9	1.39	1.11	1.14	1.12	1.08	0.99	1.65	1.43	1.40	1.52
	1:49	6.31	1.79	1.65	2.43	1.56	0.97	6.16	2.33	1.79	4.54
	1:99	13.48	2.42	2.13	3.95	2.05	1.02	12.77	3.23	2.17	8.89
	1:199	28.84	3.42	2.89	6.79	2.54	1.01	26.55	4.38	2.72	17.95
	1:399	61.28	4.86	3.79	11.71	2.84	0.95	55.65	5.75	3.37	36.76
2.5×10^{-6}	1:1	1.24	1.03	1.54	1.11	0.94	1.03	1.38	1.19	1.38	1.40
	1:9	2.47	1.19	1.45	1.29	1.41	0.77	2.51	1.61	1.49	2.23
	1:49	28.27	1.80	1.91	6.88	2.91	1.06	23.70	3.24	1.98	16.69
	1:99	89.53	2.67	1.81	16.34	3.80	0.90	71.32	4.87	1.60	42.59
	1:199	262.60	4.62	3.16	39.57	5.99	1.17	211.90	7.93	2.80	126.26
	1:399	715.05	6.70	3.90	86.87	6.65	0.80	577.70	10.70	2.83	347.10

Table S2.2. Type I error rate divided by α of robust SKAT-O when testing an association with dichotomous traits at stringent α levels $\alpha = 10^{-2}, 10^{-4}$ and 2.5×10^{-6} with three different region length (1) 1kb; (2) 2kb; (3) 3kb. The sample size was 50,000 and 10^7 datasets were generated.

α	Case: control	Region length 1kb Mean:16.33 SD: 4.05	Region length 2kb Mean:32.69 SD: 5.65	Region length 3kb Mean:49.05 SD: 6.71
10^{-2}	1:1	1.11	1.11	1.10
	1:9	1.13	1.11	1.11
	1:49	1.23	1.19	1.18
	1:99	1.33	1.29	1.27
10^{-4}	1:1	1.32	1.15	1.14
	1:9	1.40	1.18	1.20
	1:49	1.79	1.73	1.67
	1:99	2.17	2.09	1.99
2.5×10^{-6}	1:1	1.38	0.96	1.04
	1:9	1.49	1.56	1.08
	1:49	1.98	2.24	1.60
	1:99	1.60	2.44	2.08

Table S2.3. Type I error rate divided by α of different methods when testing an association between all variants, including both common and rare variants, and dichotomous traits at stringent α levels $\alpha = 10^{-2}, 10^{-4}$ and 2.5×10^{-6}. The sample size was 50,000 and 10^7 datasets were generated.

α	Case: control	SKAT-CommonRare	Robust SKAT-CommonRare without additional adjustment	Robust SKAT-CommonRare
10^{-2}	1:1	1.00	0.99	0.99
	1:9	1.00	1.01	1.01
	1:49	1.11	1.22	1.22
	1:99	1.26	1.42	1.41
	1:199	1.52	1.71	1.72
	1:399	1.97	2.15	2.15
10^{-4}	1:1	0.98	0.98	1.04
	1:9	1.09	1.11	1.13
	1:49	2.42	1.73	1.69
	1:99	4.20	2.26	2.21
	1:199	7.84	3.02	3.03
	1:399	16.02	4.01	4.04
2.5×10^{-6}	1:1	0.94	1.05	1.66
	1:9	1.53	1.17	1.45
	1:49	7.48	1.57	1.94
	1:99	18.54	2.50	1.93
	1:199	50.43	3.87	3.37
	1:399	140.09	4.47	4.46

Table S2.4. Gene-phenotype associations detected by Robust SKAT-O across 791 phenotypes at $\alpha=2.5\times10^{-6}$ (Number of associations=111)

Phenotype	Gene Name	Case	Control	NSNP	Total minor counts for cases	Total minor counts for controls	Robust SKAT	Robust burden	Robust SKAT-O
Myeloproliferative disease	*JAK2*	94	9306	73	27	442	1.94E-34	4.22E-14	1.36E-33
Unspecified monoarthritis	*OGG1*	1728	41060	117	118	1643	1.73E-06	2.33E-07	7.73E-09
Menopausal and postmenopausal disorders	*NFE2L3*	1345	21226	171	145	1358	3.80E-06	9.95E-08	2.54E-08
Cancer of prostate	*HOXB13*	741	18940	37	18	154	6.90E-08	2.15E-05	3.00E-08
Other aneurysm	*P3H1*	164	16236	110	17	497	3.45E-07	1.09E-05	5.76E-08
Heartburn	*USP45*	189	18711	103	24	649	2.04E-05	2.06E-07	6.34E-08
Fracture of hand or wrist	*GSDMC*	382	37818	109	25	761	1.72E-05	6.92E-07	7.12E-08
Congenital coagulation defects	*F11*	76	7524	38	8	84	1.96E-06	2.00E-06	7.40E-08
Congenital anomalies of great vessels	*SLC46A1*	134	13266	28	11	255	9.08E-08	3.88E-05	9.38E-08
Peptic ulcer (excl. esophageal)	*LMNB2*	773	44818	171	24	508	5.01E-07	7.63E-06	9.89E-08
Spondylosis and allied disorders	*MAP3K7CL*	849	43787	59	14	155	1.99E-07	9.40E-06	1.20E-07
Large cell lymphoma	*TNC*	56	5544	132	14	482	9.60E-08	3.48E-04	1.30E-07
Generalized convulsive epilepsy	*KRT6A*	52	5148	49	16	318	1.62E-07	1.61E-05	1.43E-07
Abnormal findings on exam of gastrointestinal tract/ abdominal area	*OR51T1*	169	16731	45	13	204	8.77E-05	1.91E-07	1.46E-07
Other open wound of head and face	*CRACR2B*	304	30096	108	9	380	1.23E-07	1.70E-02	1.51E-07
Cancer of urinary organs (incl. kidney and bladder)	*ATP6V1H*	421	41679	70	13	656	1.12E-07	1.54E-02	1.62E-07
Myeloproliferative disease	*CTAGE4;CTAGE8*	94	9306	90	25	1667	5.86E-08	7.23E-02	2.17E-07
Agorophobia, social phobia, and panic disorder	*GPR182*	78	7722	34	6	69	6.22E-07	3.65E-05	2.98E-07
Cervical cancer	*OR1J4*	159	15741	33	9	136	5.31E-07	7.89E-06	3.04E-07
Other disorders of metabolism	*ZNF746*	84	8316	52	6	164	1.37E-07	4.99E-03	3.07E-07
Other and unspecified disc disorder	*ODF2*	447	43796	201	48	2294	5.72E-07	5.50E-06	4.12E-07
Polyp of corpus uteri	*CD300E*	1059	23305	30	14	45	2.14E-04	3.07E-07	4.32E-07
Erythematous conditions	*FAM129C*	267	26433	127	24	676	1.50E-04	2.15E-07	4.44E-07
Disorders of penis	*RINL*	360	18836	74	12	163	3.07E-07	4.42E-05	5.15E-07
Noninflammatory female genital disorders	*STX17*	1139	23792	36	39	335	1.32E-03	5.15E-07	5.49E-07
Benign neoplasm of breast	*C8orf33*	177	17523	45	15	336	4.45E-07	9.12E-06	5.70E-07

Spondylosis without myelopathy	*RASA4;RASA4B*	553	43796	57	27	933	8.10E-07	8.67E-05	6.47E-07
Asthma	*MCL1*	4256	40751	84	63	301	1.51E-04	7.61E-07	6.50E-07
Rheumatoid arthritis	*ZNF462*	493	41022	435	73	4453	1.45E-07	1.40E-02	6.92E-07
Epilepsy, recurrent seizures, convulsions	*MAK*	497	44306	116	21	731	6.79E-07	1.74E-04	7.17E-07
Other disorders of male genital organs	*ARIH1*	325	18835	36	9	115	9.92E-07	1.35E-04	7.24E-07
Open wounds of head; neck; and trunk	*CRACR2B*	394	39006	134	9	502	3.33E-07	9.65E-02	7.41E-07
Scar conditions and fibrosis of skin	*C1orf109*	265	26235	54	10	242	8.86E-07	1.74E-04	7.53E-07
Diseases of the oral soft tissues, excluding lesions specific for gingiva and tongue	*ECH1*	533	44851	82	14	241	2.27E-05	1.24E-06	7.90E-07
Disturbance of skin sensation	*CXCR5*	289	28611	55	7	147	8.33E-07	4.52E-04	7.98E-07
Redundant prepuce and phimosis/BXO	*RINL*	296	18837	73	9	163	3.50E-07	5.58E-04	8.22E-07
Cardiac congenital anomalies	*SLC46A1*	207	20493	42	13	413	5.69E-07	1.90E-04	9.16E-07
Cardiomyopathy	*STXBP2*	116	11484	71	11	234	3.18E-06	1.61E-05	9.20E-07
Hearing loss	*REEP3*	372	36828	40	7	99	1.62E-06	3.12E-05	9.45E-07
Hypertension complicating pregnancy, childbirth, and the puerperium	*CLDN18*	173	17127	36	8	271	4.29E-07	1.21E-02	9.49E-07
Retention of urine	*RNF17*	722	42994	222	63	2237	4.61E-07	3.37E-04	9.59E-07
Hyposmolality and/or hyponatremia	*DLX3*	144	14256	26	12	179	2.39E-04	6.68E-07	1.06E-06
Malignant neoplasm of female breast	*ST6GALNAC6*	1307	22672	63	24	177	1.48E-06	1.64E-04	1.07E-06
Irregular menstrual cycle	*GTPBP2*	219	21243	39	6	44	6.44E-06	4.45E-06	1.07E-06
Hypertensive heart and/or renal disease	*FKBP15*	181	17919	151	13	367	1.34E-06	9.22E-05	1.08E-06
Coagulation defects	*F11*	125	12375	56	9	150	2.04E-06	1.90E-05	1.10E-06
Intracerebral hemorrhage	*GPR108*	66	6534	50	20	381	2.42E-06	2.29E-06	1.11E-06
Nasal polyps	*MPO*	456	43262	208	51	2313	2.39E-05	2.00E-06	1.15E-06
Osteoarthrosis, localized, primary	*MRPS14*	1040	42647	49	12	118	2.90E-06	2.64E-05	1.15E-06
Primary/intrinsic cardiomyopathies	*STXBP2*	114	11286	71	11	232	3.34E-06	1.50E-05	1.16E-06
Regional enteritis	*FN3KRP*	189	18711	40	19	495	1.70E-05	1.35E-06	1.19E-06
Epistaxis or throat hemorrhage	*LIN7B*	267	26433	37	6	103	1.25E-06	6.50E-04	1.20E-06
Cancer, suspected or other	*PCDHGB6*	1683	41603	165	138	2379	7.19E-07	2.25E-04	1.20E-06
Other symptoms/disorders or the urinary system	*OR10G3*	2126	42946	65	87	992	9.50E-06	1.48E-06	1.21E-06
Fever of unknown origin	*HOXB5*	455	45133	41	6	96	8.54E-07	3.40E-04	1.21E-06
Infection/inflammation of internal prosthetic device; implant; and graft	*UFSP1*	236	23364	40	12	209	2.50E-05	1.29E-06	1.22E-06
Obstruction of bile duct	*CNST*	78	7722	40	11	160	4.48E-04	4.25E-07	1.23E-06

Other local infections of skin and subcutaneous tissue	*OR4C6*	144	14256	41	9	162	2.34E-06	3.70E-05	1.26E-06
Cough	*IDH3A*	337	33363	40	30	1043	8.38E-06	5.14E-07	1.31E-06
Malignant neoplasm of uterus	*DTD1*	126	12474	15	4	63	1.08E-06	1.81E-03	1.32E-06
Type 2 diabetes with neurological manifestations	*SCAF8*	58	5742	75	10	214	1.72E-06	1.01E-04	1.37E-06
Polymyalgia Rheumatica	*RPL3*	97	9603	35	6	43	1.72E-05	1.12E-06	1.37E-06
Noninflammatory female genital disorders	*ACAP1*	1139	23792	101	19	216	3.39E-07	2.31E-02	1.39E-06
Cervical intraepithelial neoplasia [CIN] [Cervical dysplasia]	*NAGS*	309	21875	78	18	334	1.17E-04	1.53E-06	1.41E-06
Intracranial hemorrhage	*OAS1*	172	17028	66	6	172	4.79E-07	4.15E-03	1.43E-06
Cardiac and circulatory congenital anomalies	*SLC46A1*	219	21681	44	13	441	9.10E-07	4.17E-04	1.44E-06
Postmenopausal bleeding	*NFE2L3*	1171	21228	170	125	1358	9.47E-05	9.25E-07	1.44E-06
Gram negative septicemia	*SLC26A3*	87	8613	63	8	100	1.52E-06	1.34E-05	1.45E-06
Nonspecific abnormal findings in stool contents	*SFMBT1*	184	18216	82	9	163	4.51E-06	3.13E-05	1.48E-06
Acute pancreatitis	*LONRF1*	190	18810	78	13	238	6.80E-04	1.03E-06	1.49E-06
Electrolyte imbalance	*DLX3*	345	34155	39	17	417	2.31E-03	1.39E-06	1.53E-06
Other open wound of head and face	*LCTL*	304	30096	98	15	927	5.72E-07	6.68E-02	1.59E-06
Edema	*SFPQ*	160	15840	65	6	201	1.00E-06	2.91E-02	1.59E-06
Breast cancer	*ST6GALNAC6*	1423	43066	106	26	376	3.88E-06	9.41E-05	1.70E-06
Other specified benign mammary dysplasias	*EVPLL*	82	8118	39	6	57	2.28E-06	2.61E-05	1.70E-06
Symptoms involving digestive system	*ANKRD35*	2001	37239	234	143	1786	1.55E-05	4.91E-06	1.71E-06
Psoriatic arthropathy	*OR9Q1*	72	7128	22	11	183	2.34E-05	1.07E-06	1.74E-06
Rheumatoid arthritis and other inflammatory polyarthropathies	*PNOC*	582	41021	35	9	108	2.01E-06	6.07E-05	1.77E-06
Fracture of ankle and foot	*WNT11*	227	22473	69	5	152	5.76E-07	1.59E-02	1.81E-06
Open wounds of extremities	*DICER1*	435	43065	227	38	1574	9.60E-05	1.15E-06	1.84E-06
Cancer, suspected or other	*PCDHGC5*	1683	41603	179	50	813	9.81E-07	7.36E-03	1.85E-06
Respiratory abnormalities	*CBR4*	72	7128	28	8	192	1.51E-06	1.56E-03	1.90E-06
Other disorders of gallbladder	*OR13F1*	141	13959	30	7	109	1.97E-06	8.61E-05	1.91E-06
Hypertensive chronic kidney disease	*FKBP15*	150	14850	137	11	304	2.92E-06	1.87E-04	1.95E-06
Phlebitis and thrombophlebitis of lower extremities	*RELL1*	463	41305	49	10	221	1.51E-06	3.88E-04	1.98E-06
Convulsions	*TMEM5*	196	19404	47	12	225	1.17E-05	3.27E-06	2.01E-06
Other specified benign mammary dysplasias	*TRAIP*	82	8118	39	6	94	2.16E-06	4.19E-04	2.02E-06

Inflammatory bowel disease and other gastroenteritis and colitis	*GNA15*	476	37276	77	26	734	3.58E-06	7.82E-06	2.03E-06
Vascular insufficiency of intestine	*COL5A3*	53	5247	113	10	228	3.72E-06	1.47E-04	2.05E-06
Benign neoplasm of colon	*CCDC47*	2257	43090	71	40	457	2.25E-06	4.36E-03	2.09E-06
Breast cancer [female]	*ST6GALNAC6*	1398	22668	63	25	177	3.35E-06	1.68E-04	2.11E-06
Viral hepatitis	*SLC28A1*	73	7227	55	6	140	9.76E-07	2.01E-03	2.15E-06
Hemorrhoids	*GLS*	2564	41248	69	25	187	1.87E-06	8.14E-04	2.18E-06
Chronic glomerulonephritis, NOS	*ADGRL1*	75	7425	102	11	239	8.03E-06	1.56E-05	2.20E-06
Atherosclerosis of the extremities	*DONSON*	67	6633	42	9	199	2.37E-06	2.50E-03	2.22E-06
Cough	*C20orf173*	337	33363	31	5	47	2.59E-06	1.01E-04	2.22E-06
Poisoning by anticonvulsants and anti-Parkinsonism drugs	*ENTPD6*	60	5940	50	8	172	2.47E-06	1.79E-04	2.23E-06
Cerebral ischemia	*CHTOP*	341	33759	45	7	141	1.83E-06	3.86E-04	2.24E-06
Secondary malignant neoplasm	*ARNTL*	784	41638	65	10	112	7.98E-06	3.32E-05	2.28E-06
Elevated blood pressure reading without diagnosis of hypertension	*FKBP5*	167	16533	43	9	178	2.45E-06	1.02E-04	2.29E-06
Urinary incontinence	*FAM98C*	1012	42980	95	27	503	3.06E-06	1.13E-04	2.30E-06
Renal colic	*SLC10A1*	279	27621	84	11	264	5.60E-06	7.70E-05	2.30E-06
Spondylosis without myelopathy	*MAP3K7CL*	553	43796	58	9	155	1.91E-06	2.57E-04	2.32E-06
Duodenal ulcer	*RBFOX3*	322	31878	59	8	205	1.58E-06	8.63E-04	2.34E-06
Disorders of fluid, electrolyte, and acid-base balance	*DDX42*	675	44906	121	38	1304	1.80E-06	2.79E-04	2.35E-06
Cholelithiasis with acute cholecystitis	*FAM111A*	134	13266	65	16	356	8.95E-06	1.46E-05	2.39E-06
Nontoxic uninodular goiter	*EIF4G1*	55	5445	110	12	708	6.39E-07	5.54E-02	2.40E-06
Fracture of radius and ulna	*SVOP*	508	43549	61	10	176	1.40E-05	9.32E-06	2.41E-06
Other and unspecified disorders of back	*PRKAG2*	221	21879	73	10	149	1.32E-04	2.84E-06	2.47E-06
Cancer of urinary organs (incl. kidney and bladder)	*SMYD2*	421	41679	75	15	532	1.38E-06	3.25E-04	2.48E-06
Memory loss	*PRDM6*	86	8514	65	16	581	2.75E-06	3.74E-04	2.49E-06

Table S2.5. Top 3 single rare-variant signals of associations with p-value < 10^{-7} in the UK Biobank WES data.

Phenotype (Phecode)	Gene Name	RS ID	Location	MAF	P-value	Annotation	Polyphen	SIFT
Myeloproliferative disease (200)	*JAK2*	rs77375493	9:5073770:G:T	1.38E-03	1.81E-41	nonsynonymous SNV	probably_damaging	deleterious
		-	9:5022213:G:T	1.06E-04	7.16E-03	nonsynonymous SNV	probably_damaging	deleterious
		rs150221602	9:5081828:G:C	1.12E-03	5.07E-02	nonsynonymous SNV	benign	deleterious
Unspecified monoarthritis (716.2)	*OGG1*	rs113561019	3:9756791:G:A	5.67E-03	4.67E-04	nonsynonymous SNV	probably_damaging	deleterious
		rs17050550	3:9751060:G:T	2.66E-03	5.96E-04	nonsynonymous SNV	benign	tolerated
		rs104893751	3:9750423:G:A	3.89E-03	1.69E-03	nonsynonymous SNV	probably_damaging	deleterious
Menopausal and postmenopausal disorders (627)	*NFE2L3*	rs148235978	7:26184630:A:G	7.93E-03	2.72E-05	nonsynonymous SNV	benign	deleterious
		rs148159120	7:26185140:G:A	5.91E-03	2.18E-03	nonsynonymous SNV	benign	tolerated
		rs144789579	7:26185559:A:G	3.33E-04	6.71E-03	nonsynonymous SNV	benign	tolerated
Cancer of prostate (185)	*HOXB13*	rs138213197	17:48728343:C:T	2.16E-03	5.24E-08	nonsynonymous SNV	probably_damaging	deleterious
		rs764401781	17:48728250:G:A	7.62E-05	3.37E-02	nonsynonymous SNV	benign	deleterious
		rs774579054	17:48728379:C:A	1.02E-04	7.05E-02	nonsynonymous SNV	possibly_damaging	tolerated
Other aneurysm (442)	*P3H1*	rs372710498	1:42766778:C:T	6.71E-04	1.71E-05	nonsynonymous SNV	probably_damaging	tolerated
		rs140254470	1:42748243:C:T	3.35E-04	1.21E-04	nonsynonymous SNV	benign	tolerated
		rs372301077	1:42758970:C:G	9.15E-05	3.88E-04	nonsynonymous SNV	probably_damaging	deleterious
Heartburn (530.9)	*USP45*	rs554927779	6:99468543:D:1	4.52E-03	5.39E-05	frameshift deletion	LoF: High-confidence	-
		rs201110065	6:99508768:G:T	3.70E-04	2.13E-03	nonsynonymous SNV	probably_damaging	deleterious_low_confidence

		rs144269307	6:99488237:G:A	2.65E-05	2.78E-03	nonsynonymous SNV	possibly_damaging	tolerated
Fracture of hand or wrist (804)	*GSDMC*	rs149748731	8:129748706:G:A	3.17E-03	8.17E-05	nonsynonymous SNV	possibly_damaging	tolerated
		rs79403769	8:129776220:C:A	1.14E-03	4.95E-03	nonsynonymous SNV	benign	deleterious
		rs16904151	8:129765749:C:T	6.54E-04	5.81E-03	nonsynonymous SNV	benign	tolerated
Congenital coagulation defects (286.1)	*F11*	rs281875276	4:186288589:T:G	1.32E-04	4.52E-05	nonsynonymous SNV	probably_damaging	deleterious
		rs140190776	4:186273178:G:A	1.97E-04	1.73E-04	splicing	LoF: High-confidence	-
		-	4:186267139:G:T	6.58E-05	3.85E-03	nonsynonymous SNV	probably damaging	deleterious
Congenital anomalies of great vessels (747.13)	*SLC46A1*	rs189103810	17:28405185:A:T	1.75E-03	1.86E-08	nonsynonymous SNV	possibly_damaging	deleterious
		rs281875211	17:28402276:C:T	3.73E-05	1.24E-02	nonsynonymous SNV	possibly_damaging	deleterious
		rs41297071	17:28405926:C:G	5.22E-04	7.13E-01	nonsynonymous SNV	benign	tolerated
Peptic ulcer (excl. esophageal) (531)	*LMNB2*	rs752957340	19:2434094:G:A	2.30E-04	3.83E-06	nonsynonymous SNV	probably_damaging	deleterious
		-	19:2456926:G:T	2.19E-05	1.47E-04	nonsynonymous SNV	-	-
		-	19:2431845:C:A	1.10E-05	3.45E-03	nonsynonymous SNV	benign	deleterious

Table S2.6. The most significant nearby variant association signals (± 100 Kbp up and down stream) in the UK-Biobank imputed datasets of 400,000 British samples.

Phenotype (Phecode)	Gene Name	RS ID	Location	Ref Allele	Alt Allele	MAF	p-value
Myeloproliferative disease (200)	*JAK2*	rs10283564	chr9:5075628	C	G	2.51E-01	2.30E-17
Unspecified monoarthritis (716.2)	*OGG1*	rs75924392	chr3:9797369	A	G	4.25E-02	4.28E-04
Menopausal and postmenopausal disorders (627)	*NFE2L3*	rs35838658	chr7:26175486	T	A	3.48E-01	2.14E-04
Cancer of prostate (185)	*HOXB13*	rs145922598	chr17:48733224	C	T	3.27E-02	1.17E-04
Other aneurysm (442)	*P3H1*	rs148797090	chr1:42818294	T	C	1.35E-02	1.22E-03
Heartburn (530.9)	*USP45*	rs75597991	chr6:99367094	G	C	2.38E-02	4.08E-02
Fracture of hand or wrist (804)	*GSDMC*	rs409790	chr8:129833249	A	G	8.28E-01	1.84E-02
Congenital coagulation defects (286.1)	*F11*	rs115583874	chr4:186260648	G	T	1.37E-02	6.30E-03
Congenital anomalies of great vessels (747.13)	*SLC46A1*	rs111306228	chr17:28425402	C	A	2.06E-02	2.29E-03
Peptic ulcer (excl. esophageal) (531)	*LMNB2*	rs146283067	chr19:2441769	G	T	1.62E-02	1.31E-03

CHAPTER III

Scalable Generalized Linear Mixed Model for Region-based Association Tests in Large Biobanks and Cohorts

Abstract

With very large sample sizes, population-based cohorts and biobanks provide an exciting opportunity to identify genetic components of complex traits. To analyze rare variants, gene or region-based multiple variant aggregate tests are commonly used to increase association test power. However, due to the substantial computation cost, existing region-based rare variant tests cannot analyze hundreds of thousands of samples while accounting for confounders, such as population stratification and sample relatedness. Here we propose a scalable generalized mixed model region-based association test that can handle large sample sizes and accounts for unbalanced case-control ratios for binary traits. This method, SAIGE-GENE, utilizes state-of-the-art optimization strategies to reduce computational and memory cost, and hence is applicable to exome-wide and genome-wide region-based analysis for hundreds of thousands of samples. Through the analysis of the HUNT study of 69,716 Norwegian samples and the UK Biobank data of 408,910 White British samples, we show that SAIGE-GENE can efficiently analyze large sample data (N > 400,000) with type I error rates well controlled.

3.1 Introduction

In recent years, large cohort studies and biobanks, such as Trans-Omics for Precision Medicine (TOPMed) study(Taliun et al., 2019) and UK Biobank(C. Bycroft et al., 2018), have sequenced or genotyped hundreds of thousands of samples, which are invaluable resources to identify genetic components of complex traits, including rare variants (minor allele frequency (MAF) < 1%). It is well known that single variant tests are underpowered to identify trait-associated rare variants(S. Lee et al., 2014). Gene- or region-based tests, such as Burden test, SKAT(M. C. Wu et al., 2011) and SKAT-O(S. Lee et al., 2012), can be more powerful by grouping rare variants into functional units, i.e. genes. To adjust for both population structure and sample relatedness, gene-based tests have been extended to mixed models(H. Chen et al., 2019). For example, EmmaX(Kang et al., 2010) based SKAT(M. C. Wu et al., 2011) approaches (EmmaX-SKAT) have been implemented and used for many rare variant association studies including TOPMed(Natarajan et al., 2018; Taliun et al., 2019). The generalized linear mixed model gene-based test, SMMAT, has been recently developed(H. Chen et al., 2019). However, these approaches require O(N^3) computation time and O(N^2) memory usages, where N is the sample size, which are not scalable to large datasets.

Here, we propose a novel method called SAIGE-GENE for region-based association analysis that is capable of handling very large samples (> 400,000 individuals), while inferring and accounting for sample relatedness. SAIGE-GENE is an extension of the previously developed single variant association method, SAIGE(W. Zhou et al., 2018), with a modification suitable to rare variants. Same as SAIGE, it utilizes state-of-the-art optimization strategies to reduce computation cost for fitting null mixed models. To ensure computation efficiency while improving test accuracy for rare variants, SAIGE-GENE approximates the variance of score statistics calculated with the full

genetic relationship matrix (GRM) using the variance calculated with a sparse GRM and the ratios of these two variances estimated from a subset of genetic markers. Because the sparse GRM, which is constructed by thresholding small values in the full GRM, preserves close family structures, this approach provides a far more accurate variance estimation for very rare variants (minor allele count (MAC) < 20) than the original approach in SAIGE(W. Zhou et al., 2018). By combining single variant score statistics, SAIGE-GENE can perform Burden, SKAT and SKAT-O type gene-based tests. We have also developed conditional analysis to perform association tests conditioning on a single variant or multiple variants to identify independent rare variant association signals. Furthermore, SAIGE-GENE can account for unbalanced case-control ratios of binary traits by adopting a robust adjustment based on saddlepoint approximation(Daniels, 1954; R. Dey et al., 2017; Kuonen, 1999) (SPA) and efficient resampling(Lee et al., 2016) (ER). The robust adjustment was previously developed for independent samples(Zhao et al., 2019) and we have extended it for related samples in SAIGE-GENE.

We have demonstrated that SAIGE-GENE controls for type I error rates in related samples for both quantitative and binary traits through extensive simulations as well as real data analysis, including the HUNT study for 69,716 Norwegian samples(Krokstad et al., 2013; Langhammer et al., 2012) and the UK Biobank for 408,910 White British samples(C. Bycroft et al., 2018). By evaluating its computation performance of SAIGE-GENE, we have shown its feasibility for large-scale genome-wide analysis. To perform exome-wide gene-based tests on 400,000 samples with on average 50 markers per gene, SAIGE-GENE requires 2,238 CPU hours and less than 36 Gb memory, while current methods will cost more than > 10 Tb in memory. We have further applied SAIGE-GENE to 53 quantitative traits and 10 binary traits in the UK Biobank and identified several significantly associated genes through exome-wide gene-based tests.

3.2 Methods

3.2.1 Overview of methods

SAIGE-GENE consists of two main steps: 1. Fitting the null generalized linear mixed model (GLMM) to estimate variance components and other model parameters. 2. Testing for association between each genetic variant set, such as a gene or a region, and the phenotype. Three different association tests: Burden, SKAT, and SKAT-O have been implemented in SAIGE-GENE. The workflow is shown in the **Figure S3.1**.

SAIGE-GENE uses similar optimization strategies as utilized in the original SAIGE to achieve the scalability for fitting the null GLMM and estimating the model parameters in Step 1. In particular, the spectral decomposition has been replaced by the preconditioning conjugate gradient (PCG) to solve linear systems without calculating and inverting the $N \times N$ GRM. To reduce the memory usage, raw genotypes are stored in a binary vector and elements of GRM are calculated when needed rather than being stored.

One of the most time-consuming part in association tests is to calculate variance of single variant score statistic, which requires O(N^2) computation. To reduce computation cost, existing approaches, such as SAIGE(W. Zhou et al., 2018), BOLT-LMM(Loh et al., 2015), and GRAMMA-Gamma(Svishcheva et al., 2012), approximate the variance of single variant score statistics with the full GRM using the variance estimate without a GRM and the ratio of these two variances. The ratio, which is assumed to be constant, is estimated using a subset of randomly selected genetic markers. However, for very rare variants with MAC below 20, the constant ratio assumption is not satisfied (**Figure S3.2, left panel**). This is because rare variants are more susceptible to close family structures. Thus, to better approximate the variance, SAIGE-GENE

incorporates close family structures through a sparse GRM, in which GRM elements below a user-specified relatedness coefficient are zeroed out and close family structures are preserved. The ratio between the variance with the full GRM and with the sparse GRM is much less variable (**Figure S3.2, right panel**). To construct a sparse GRM, a small subset of randomly selected genetic markers, i.e. 2,000, are firstly used to quickly estimate which sample pairs pass the user-specified coefficient of relatedness cutoff, e.g. $\geq$0.125 for up to 3rd degree relatives. Then the coefficients of relatedness for those related pairs are further estimated using the full set of genetic markers, which equal to values in the full GRM. Given that estimated values for variance ratios vary by MAC for the extremely rare variants (**Figure S3.2, left panel**), such as singletons and doubletons, the variance ratios need to be estimated separately for different MAC categories. By default, MAC categories are set to be MAC equals to 1, 2, 3, 4, 5, 6 to 10, 11 to 20, and > 20.

In Step 2, gene-based tests are conducted using single variant score statistics and their covariance estimates, which are approximated as the product of the covariance with the sparse GRM and the pre-estimated ratio. SAIGE-GENE can carry out Burden, SKAT, and SKAT-O approaches. Since SKAT-O is a combined test of Burden and SKAT, and hence provides a robust power, SAIGE-GENE performs SKAT-O by default.

If a gene or a region is significantly associated with the phenotype of interest, it is necessary to test if the signal is from rare variants or just a shadow of common variants in the same locus. We have developed conditional analysis using linkage disequilibrium (LD) information between conditioning markers and the tested gene(Liu et al., 2014). Details are described in the Online Methods section.

SAIGE-GENE uses the same generalized linear mixed model as in SMMAT, while SMMAT calculates the variances of the score statistics for all tested genes using the full GRM directly and

hence can be thought of as the "exact" method. When the trait is continuous, GLMM used by SAIGE-GENE and SMMAT is equivalent to the linear mixed model (LMM) of EmmaX-SKAT. We have further shown that SAIGE-GENE provides consistent association p-values to the two "exact" methods, EmmaX-SKAT and SMMAT (r^2 of -$\log_{10}$ p-values > 0.99), using both simulation studies (**Figure S3.3**) and real data analysis for down-sampled UK Biobank and HUNT (**Figure S3.4**), but with much smaller computation and memory cost (**Figure 3.1**). We have also shown that SAIGE-GENE with different coefficient of relatedness cutoffs (0.125 and 0.2) produced nearly identical association p-values for automated read pulse rates in UK Biobank (**Figure S3.5**).

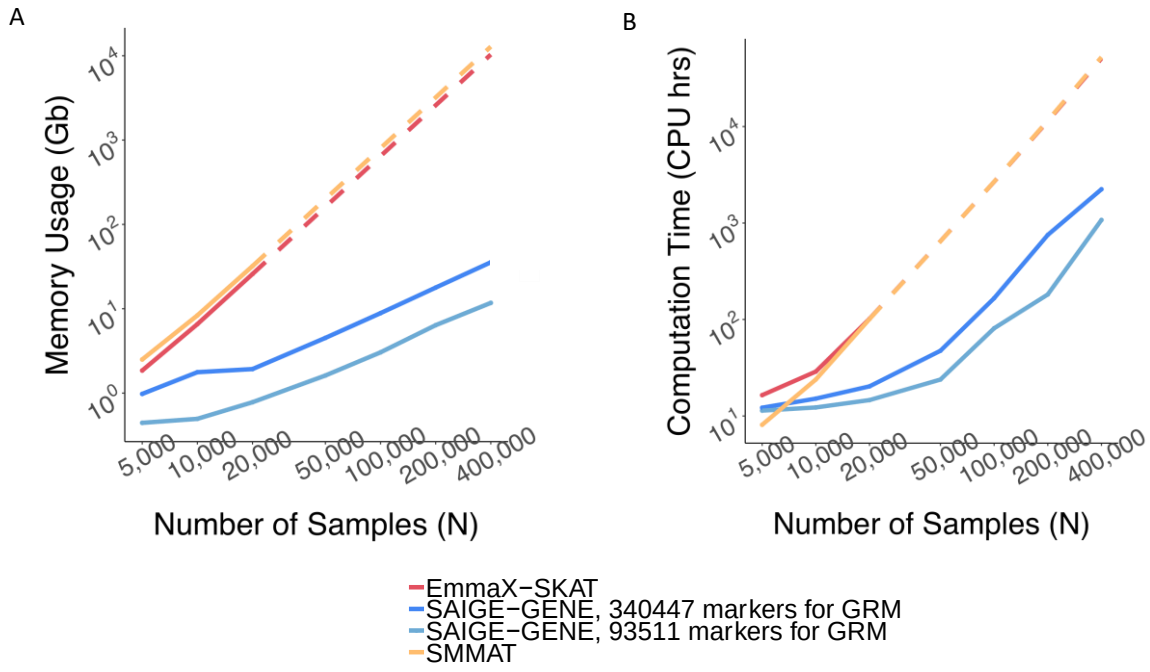


Figure 3.1 Estimated and projected computation cost by sample sizes (N) for gene-based tests for 15,342 genes, each containing 50 rare variants.
Benchmarking was performed on randomly sub-sampled UK Biobank data with 408,144 White British participants for waist-to-hip ratio. The reported run times and memory are medians of five runs with samples randomly selected from the full sample set using different sampling seeds. The reported computation time and memory for EmmaX-SKAT and SMMAT is the projected computation time when N > 20,000. A. Log-log plots of the memory usage as a function of sample size (N) B. Log-log plots of the run time as a function of sample size (N). Numerical data are provided in **Table S3.1**.

For binary phenotypes with unbalanced case-control ratios (< 1:9), single variant score statistics do not follow the normal distribution, leading to inflated type I error rates for region-based test(Lee et al., 2016). To address this problem, we have recently developed a scalable robust adjustment for independent samples(Zhao et al., 2019). The approach uses saddlepoint approximation(Daniels, 1954; R. Dey et al., 2017; Kuonen, 1999) (SPA) and efficient resampling(Lee et al., 2016) (ER) to calibrate the variance of single variant score statistics. We have extended this approach to GLMM for SAIGE-GENE, which provides greatly improved type I error control than the unadjusted approach of assuming normality (**Figure S3.6).** Details can be found in **Supplementary Materials 1.3.3**.

3.2.2 Generalized linear mixed model

In a study with sample size N, we denote the phenotype of the *ith* individual using y_i for both continuous and binary traits. Let the $1 \times (p+1)$ vector X_i represent p covariates including the intercept, the $N \times q$ matrix G_i represent the allele counts $(0, 1 \text{ or } 2)$ for q variants in the gene to test. The generalized linear mixed model can be written as

$$g(\mu_i) = X_i\alpha + G_i\beta + b_i,$$

where μ_i is the mean of phenotype, b_i is the random effect, which is assumed to be distributed as $N(0, \tau\,\psi)$, where ψ is an $N \times N$ genetic relationship matrix (GRM) and τ is the additive genetic variance parameter. The link function g is the identity function for continuous traits with an error term $\varepsilon \sim N(0, \phi I)$ and logistic function for binary traits. The parameter α is a $(p+1) \times 1$ coefficient vector of fixed effects and β is a $q \times 1$ coefficient vector of the genetic effect.

3.2.3 Estimate variance component and other model parameters (Step 1)

Same as in the original SAIGE(W. Zhou et al., 2018) and GMMAT(H. Chen et al., 2016), to fit the null GLMM in SAIGE-GENE, penalized quasi-likelihood (PQL) method(Breslow & Clayton,

1993; Lee & van der Werf, 2006) and the computational efficient average information restricted maximum likelihood (AI-REML) algorithm(H. Chen et al., 2016; Gilmour et al., 1995) are used to iteratively estimate $(\hat{\tau}, \hat{\alpha}, \hat{b})$ under the null hypothesis of $\beta = 0$. At iteration k, let $(\hat{\tau}^{(k)}, \hat{\alpha}^{(k)}, \hat{b}^{(k)})$ be estimated $(\hat{\tau}, \hat{\alpha}, \hat{b})$, $\hat{\mu}_i^{(k)}$ be the estimated mean of y_i and $\hat{\Sigma}^{(k)} = (\widehat{W}^{(k)})^{-1} + \hat{\tau}^{(k)}\psi$ be an $N \times N$ matrix of the variance of working vector $\tilde{y}_i$, in which ψ is the $N \times N$ GRM. For continuous traits, $\widehat{W}^{(k)} = \hat{\phi}^{-1}I$ and $\tilde{y}_i = X_i\alpha^{(k)} + b_i^{(k)}$. For binary traits, $\widehat{W}^{(k)} = diag[\hat{\mu}_i^{(k)}\left(1 - \hat{\mu}_i^{(k)}\right)]$ and $\tilde{y}_i = X_i\alpha^{(k)} + b_i^{(k)} + (y_i - \hat{\mu}_i^{(k)})/\left\{\hat{\mu}_i^{(k)}\left(1 - \hat{\mu}_i^{(k)}\right)\right\}$. To obtain the log quasi-likelihood and average information at each iteration, SAIGE and SAIGE-GENE use the preconditioned conjugate gradient approach (PCG) to obtain the product of inverse of $\hat{\Sigma}^{(k)}$and any other vector by iteratively solving a linear system with $\hat{\Sigma}^{(k)}$. This approach is more computationally efficient than using Cholesky decomposition to obtain $(\hat{\Sigma}^{(k)})^{-1}$.

3.2.4 Gene-based association tests (Step 2)

Test statistics of the Burden, SKAT and SKAT-O tests for a gene can be constructed based on score statistics from the marginal model for individual variants in the gene. Suppose there are q variants in the region or gene to test. The score statistic for variant j (j=1,..,q) under H_0: $\beta_j = 0$ is $T_j = g_j{}^T(Y - \hat{\mu})$ where g_j and Y are $N \times 1$ genotype and phenotype vectors, respectively, and $\hat{\mu}$ is the estimated mean of Y under the null hypothesis.

Let $u_j$denote a threshold indicator or weight for variant j and $\mathrm{U} = \mathrm{diag}(u_1, \ldots, u_q)$ be a diagonal matrix with u_j as the jth element. Similar to the original SKAT and SKAT-O papers(S. Lee et al., 2012; M. C. Wu et al., 2011), to upweight rare variants, the default setting in SAIGE-GENE is $u_j = Beta(MAF_j, 1, 25)$, which upweight rarer variants. The Burden test statistics can be written as $Q_{Burden} = \left(\sum_{j=1}^{q} u_j T_j\right)^2$. Suppose $\tilde{G} = G - X\left(X^T\widehat{W}X\right)^{-1}X^T\widehat{W}G$ is the covariate adjusted

genotype matrix, where $G = (g_1, \dots, g_q)$ is the $N \times q$ genotype matrix of the q genetic variants, and $\hat{P} = \hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}X\left(X^T\hat{\Sigma}^{-1}X\right)^{-1}X^T\hat{\Sigma}^{-1}$ with $\hat{\Sigma} = \widehat{W}^{-1} + \hat{\tau}\psi$. Under the null hypothesis of no genetic effects, Q_{Burden} followed $\lambda_B\chi_1^2$, where $\lambda_B = J^T\mathrm{U}\tilde{G}^T\hat{P}\tilde{G}\mathrm{U}J, J$ is a $q \times 1$ vector with all elements being unity and $\chi_1^2$is a chi-squared distribution with 1 degree of freedom(S. Lee et al., 2014). The SKAT test(M. C. Wu et al., 2011) can be written as $Q_{SKAT} = \sum_{j=1}^{q} u_j^2 T_j^2$, which follows a mixture of chi-square distribution $\sum_{j=1}^{q}\lambda_{Sj}\chi_1^2$, where λ_{Sj} are the eigenvalues of $U\tilde{G}^T\hat{P}\tilde{G}U$. The SKAT-O test(S. Lee et al., 2012) uses a linear combination of the Burden and SKAT tests statistics $Q_{SKATO} = (1-\rho)Q_{SKAT} + \rho Q_{Burden}, 0 \leq \rho \leq 1$. To conduct the test, the minimum p-value from grid of ρ is calculated and the p-value of the minimum p-value is estimated through numerical integration. Following the suggestion in Lee *et al*(Lee et al., 2013), we use a grid of eight values of $\rho = (0, 0.1^2, 0.2^2, 0.3^2, 0.4^2, 0.5^2, 0.5, 1)$ to find the minimum p-value.

3.2.5 Approximate $\tilde{G}^T\hat{P}\tilde{G}$

For each gene, given $\hat{P}$, the calculation of $\tilde{G}^T\hat{P}\tilde{G}$ requires applying PCG for each variant in the gene, which can be computationally very expensive. Suppose $\tilde{g}$ represents a covariate adjusted single variant genotype vector. To reduce computation cost, an approximation approach has been used in SAIGE, BOLT-LMM(Loh et al., 2015) and GRAMMAR-GAMMAR(Svishcheva et al., 2012), in which the ratio between $\tilde{g}^T\hat{P}\tilde{g}$ and $\tilde{g}^T\tilde{g}$ is estimated by a small subset of randomly selected genetic markers. The ratio has been shown to be approximately constant for all variants. Given the estimated ratio $\hat{r} = \tilde{g}^T\hat{P}\tilde{g}\ /\tilde{g}^T\tilde{g}$, $\tilde{g}^T\hat{P}\tilde{g}$ for all other variants can be obtained as $\hat{r}\tilde{g}^T\tilde{g}$. However, the variations of the estimated $\hat{r}$ for extremely rare variants are large and including some closely related samples in the denominator helps reduce the variation of $\hat{r}$ as shown in **Figure S3.2.** Let ψ_S denote a sparse GRM that preserves close family structure and ψ_f denote a full GRM. We

estimate the ratio $\hat{r}_s = \tilde{g}^T \hat{P} \tilde{g} \,/ \tilde{g}^T \hat{P}_s \tilde{g}$, where $\hat{P}_s = \hat{\Sigma}_s^{-1} - \hat{\Sigma}_s^{-1} X \left(X^T \hat{\Sigma}_s^{-1} X \right)^{-1} X^T \hat{\Sigma}_s^{-1}$ and $\hat{\Sigma}_s = \widehat{W}^{-1} + \hat{\tau}\, \psi_s$.

In ψ_s, elements below a user-specified relatedness coefficient cutoff, i.e. > 3rd degree relatedness, are zeroed out with only close family structures being preserved. To construct ψ_s, a subset of randomly selected genetic markers, i.e. 2,000, is firstly used to quickly estimate which related samples pass the user-specified cutoff. Then the relatedness coefficients for those samples are further estimated using the full set of genetic markers, which equal to corresponding values in the ψ_f. In the model fitting using ψ_s, $\hat{\Sigma}_s^{-1} X$ and $\hat{\Sigma}_s^{-1} \tilde{g}$ need to be calculated. For this we use a sparse-LU based solve method(Davis, 2006) implemented in R. The constructed ψ_s is also used for approximating the variance of score statistics with ψ_f. For a biobank or a data set, ψ_s only needs to be constructed once and can be re-used for any phenotypes in the same date set.

SAIGE-GENE estimates variance ratios for different MAC categories. By default, MAC categories are set to be MAC equals to 1, 2, 3, 4, 5, 6 to 10, 11 to 20, and is greater than 20. Once the MAC categorical variance ratios are estimated, for each genetic marker in tested genes or regions, $\hat{r}_s$ can be obtained according to its MAC. Let $\hat{R}_s$ be a $q \times q$ diagonal matrix whose jth diagonal element is the ratio $\hat{r}_s$ for the jth marker in the gene (i.e. $\tilde{g}_j^T \hat{P} \tilde{g}_j \,/ \tilde{g}_j^T \hat{P}_s \tilde{g}_j$). For the tested gene with q markers, $\tilde{G}^T \hat{P} \tilde{G}$ can be approximated as $\hat{R}_s^{\frac{1}{2}} \tilde{G}^T \hat{P}_s \tilde{G} \hat{R}_s^{\frac{1}{2}}$ (See **Supplementary Materials** for more details).

3.2.6 Robust adjustment for $\hat{R}_s^{\frac{1}{2}} \tilde{G}^T \hat{P}_s \tilde{G} \hat{R}_s^{\frac{1}{2}}$ to account for unbalanced case-control ratios

To account for unbalanced case-control ratios of binary traits in region- or gene-based tests, we recently developed a robust adjustment for independent samples(Zhao et al., 2019). The approach

first obtains well-calibrated p-values of single variant score statistics using SPA(Daniels, 1954; R. Dey et al., 2017; Kuonen, 1999) and ER(Lee et al., 2016). SPA is a method to calculate p-values by inverting the cumulant generating function (CGF). Since CGF completely specifies the distribution, SPA can be far more accurate than using the normal distribution. However, since SPA is still an asymptotic based approach, it does not work well when variants are very rare (ex. MAC $\leq$10). For those variants, we use ER, which resamples the case-control status of only individuals carrying a minor allele and is extremely fast for very rare variants. To account for the fact that individuals can have different non-genetic risk of diseases (due to covariates), the resampling was done with the estimated disease risk μ_i. Next, variances of single variant score statistics are obtained by inverting those p-values, which are then used to calibrate the variances of region- or gene-based test statistics. We have extended the approach for related samples in SAIGE-GENE. For variants with MAC > 10, single-variant p-values are obtained by SAIGE, which basically applies SPA to GLMM. For variants with MAC $\leq$10, we use ER with GLMM estimated $\widehat{\mu_i}$, which includes the random effect to maintain the correlation structure among samples. After calculating p-values of T_j for j=1,…,q, the variance of T_j is calibrated by inverting the corresponding p-value. Then the calibrated variance is applied to $\hat{R}_S^{\frac{1}{2}}\tilde{G}^T\hat{P}_s\tilde{G}\hat{R}_S^{\frac{1}{2}}$ to compute robust p-value for the region- or gene-based test. The details can be found in **Supplementary Materials**.

3.2.7 Conditional analysis

In SAIGE-GENE, we have implemented the conditional analysis to perform gene-based tests conditioning on a given markers using the summary statistics from the unconditional gene-based tests and the linkage disequilibrium r^2 between testing and conditioning markers(Liu et al., 2014). Let G be the genotypes for a gene to be tested for association, which contains q markers, and G_2 be the genotypes for the conditioning markers, which contains q_2 markers. Let β denote a $q \times 1$

coefficient vector of the genetic effect for the gene to be tested and β_2 be a $q_2 \times 1$ coefficient vector of the genetic effect for the conditioning markers. The genotype matrix with the non-genetic covariates projected out $\tilde{G} = G - X(X^T\widehat{W}X)^{-1}X^T\widehat{W}G$ and $\tilde{G}_2 = G_2 - X(X^T\widehat{W}X)^{-1}X^T\widehat{W}G_2$. In the unconditioned association tests, the test statistics $T = \tilde{G}^T(Y - \hat{\mu})$ and $T_2 = \tilde{G}_2{}^T(Y - \hat{\mu})$. In conditional analysis, under the null hypothesis, E(T) = E($\tilde{G}^T P(\tilde{G}_2\beta_2)) = \tilde{G}^T\hat{P}\tilde{G}_2\beta_2$ and E(T_2) = E($\tilde{G}_2^T P(\tilde{G}_2\beta_2)) = \tilde{G}_2^T\hat{P}_s\tilde{G}_2\beta_2$. T and T_2 jointly follow the multivariate normal with mean (E(T), E(T_2)) and variance $S = \begin{bmatrix} \tilde{G}^T\hat{P}\tilde{G} & \tilde{G}^T\hat{P}\tilde{G}_2 \\ \tilde{G}_2^T\hat{P}\tilde{G} & \tilde{G}_2^T\hat{P}\tilde{G}_2 \end{bmatrix}$.

Thus under the null hypothesis of no association of T, i.e. H_0: $\beta = 0$, the $T|T_2$ follows the conditional normal distribution with E($T|T_2$) = $\tilde{G}^T\hat{P}\tilde{G}_2\ (\tilde{G}_2^T\hat{P}\tilde{G}_2)^{-1}T2$ and var($T|T_2$) = $\tilde{G}^T\hat{P}\tilde{G} - \tilde{G}^T\hat{P}\tilde{G}_2\ \left(\tilde{G}_2^T\hat{P}\tilde{G}_2\right)^{-1}\tilde{G}_2^T\hat{P}\tilde{G}$, and p-values can be calculated from the conditional distribution.

3.2.8 Data simulation

We carried out a series of simulations to evaluate and compare the performance of SAIGE-GENE, EmmaX-SKAT(Kang et al., 2010; S. Lee et al., 2012) and SMMAT(H. Chen et al., 2019). We used the sequence data from 10,000 European ancestry chromosomes over 1Mb regions that was generated using the calibrated coalescent model in the SKAT R package(S. Lee et al., 2012). We randomly selected 10,000 regions with 3Kb from the sequence data, followed by the gene-dropping simulation[44] using these sequences as founder haplotypes that were propagated through the pedigree of 10 family members shown in **Figure S3.11**. Only variants with MAF $\leq$ 1% were used for simulation studies. Quantitative phenotypes were generated from the following linear mixed model $y_i = X_1 + X_2 + G_i\beta + b_i + \varepsilon_i$, where G_i is the genotype value, β is the genetic effect sizes, b_i is the random effect simulated from $N(0, \tau\psi)$, and ε_i is the error term simulated from $N(0, (1-\tau)I)$. Two covariates, X_1 and X_2, were simulated from Bernoulli(0.5) and N(0,1),

respectively. Binary phenotypes were generated from the logistic mixed model $logit(\pi_{i0}) = \alpha_0 + b_i + X_1 + X_2 + G_i\beta$, where β is the genetic log odds ratio, b_i is the random effect simulated from $N(0, \tau\psi)$ with $\tau = 1$. The intercept α_0 was determined by the disease prevalence (i.e. case-control ratios). Given $\tau = 1$, the liability scale heritability is 0.23[45].

To evaluate the type I error rates at exome-wide $\alpha=2.5\times10^{-6}$, we first simulated 10,000 regions, and then simulated 1000 sets of quantitative phenotypes for each simulated region with different random seeds under the null hypothesis with $\beta = 0$. Gene-based association tests were performed using SAIGE-GENE, EmmaX-SKAT, and SMMAT therefore in total 10^7 tests for each of Burden, SKAT, and SKAT-O tests were carried out. Two different settings for τ were evaluated: 0.2 and 0.4 and two different sample relatedness settings were used: one has 500 families and 5,000 independent samples and other one has 1,000 families, each with 10 family members. We also simulated 1,000 sets of binary phenotypes for case-control ratios 1:99, 1:19, 1:9, 1:4, and 1:1 for 500 families and 5,000 independent samples. Burden, SKAT, and SKAT-O tests were performed on the 10,000 genome regions using SAIGE-GENE, in total 10^7 tests for each method for each case-control ratio.

For the power simulation, phenotypes were generated under the alternative hypothesis $\beta \neq 0$. Two different settings for proportions of causal variants are used: 10% and 40%, corresponding to $|\beta| = |log10(MAF)|$ and $|\beta| = |0.3log10(MAF)|$, respectively. In each setting, 80% and 100% had negative effect sizes. We simulated 1,000 datasets in each simulation, and power was evaluated at test-specific empirical α, which yields nominal $\alpha=2.5\times10^{-6}$. The empirical α was estimated from the type I error simulations. Similarly, 1,000 sets of binary traits were generated for 10,000 samples (500 families and 5,000 independent samples) under the alternative hypothesis $\beta \neq 0$ using two different settings: cohort study with various disease prevalence (0.01, 0.05, 0.1, and 0.5); and case-

control sampling with three different case-control ratios (1:19, 1:9, and 1:1) based on a disease prevalence 1% in the population (**Supplementary Materials 2.5**). 40% variants are simulated as causal variants, among which 80% are risk-increasing variants and 20% are risk-decreasing. The absolute effect sizes of causal variants are set to be $|0.55\log_{10}(MAF)|$ and $|0.35\log_{10}(MAF)|$ for cohort study and case-control sampling, respectively.

3.2.9 HUNT and UK Biobank data analysis

We applied SAIGE-GENE to the high-density lipoprotein (HDL) levels in 69,500 Norwegian samples from a population-based Nord Trøndelag Health Study (HUNT) [9]. About 70,000 HUNT participants were genotyped using Illumina HumanCoreExome v1.0 and 1.1 and imputed using Minimac3(Das et al., 2016) with a merged reference panel of HRC and whole genome sequencing data (WGS) for 2,201 HUNT samples. Variants with imputation $r^2 < 0.8$ were excluded from further analysis. Total 13,416 genes with at least two rare (MAF $\leq$ 1%) missense and/or stop-gain variants with imputation $r^2 \geq 0.8$ were tested. Variants were annotated using Seattle Seq Annotations (http://snp.gs.washington.edu/SeattleSeqAnnotation138/). Age, Sex, genotyping batch, and first four PCs were included as covariates in the model. We used 249,749 pruned genotyped markers to estimate relatedness coefficients in the full GRM for Step 1 and used the relative coefficient cutoff ≥ 0.125 for the sparse GRM.

We have also analyzed 53 quantitative traits and 10 binary traits using SAIGE-GENE in the UK Biobank for 408,910 participants with White British ancestry(C. Bycroft et al., 2018). Markers that were imputed by the Haplotype Reference Consortium (HRC)[20] panel with imputation info score ≥ 0.8 were used in the analysis. Total 15,342 genes with at least two rare (MAF $\leq$ 1%) missense and stop-gain variants that were directly genotyped or successfully imputed from HRC (imputation score ≥ 0.8) were tested. Sex, age when attended assessment center, and first four PCs

that were estimated using all samples with White British ancestry were adjusted in all tests. We used 340,447 pruned markers, which were pruned from the directly genotyped markers using the following parameters, were used to construct GRM: window size of 500 base pairs (bp), step-size of 50 bp, and pairwise $r^2 < 0.2$. We used the relative coefficient cutoff ≥ 0.125 for the sparse GRM.

3.3 Results

3.3.1 Computation and Memory Cost

To evaluate the computation performance of SAIGE-GENE, we randomly sampled subsets of the 408,144 UK Biobank participants with the White British ancestry and non-missing measurements for waist hip ratio(C. Bycroft et al., 2018). We benchmarked SAIGE-GENE, EmmaX-SKAT, and SMMAT for exome-wide gene-based SKAT-O tests, in which 15,342 genes were tested with assuming that each has 50 rare variants.

Memory usage is plotted on a log10 scale against sample sizes in **Figure 3.1A**. The memory cost of SAIGE-GENE is linear to the number of markers, M_1, used for kinship estimation, but using too few markers may not be sufficient to account for subtle sample relatedness in the data, leading to inflated type I error rates in genetic association tests(Yang et al., 2014; W. Zhou et al., 2018). SAIGE-GENE uses 11.74 Gb with M_1 = 93,511 and 35.59 Gb when M_1 = 340,447 when the sample size N is 400,000, making it feasible for large sample data. In contrast, with N = 400,000 the memory usages in EmmaX-SKAT and SMMAT are projected to be nearly 10Tb, which makes them impossible to be used for large sample data.

Total computation time for exome-wide gene-based tests is plotted on a log10 scale against the sample size as shown in **Figure 3.1B.** Computation time for Step 1 and Step 2 are plotted

separately in **Figure S3.7** with numbers presented in **Table S3.1**. The computation time for Step 1 in SAIGE-GENE is approximately O($M_1N^{1.5}$) and in SMMAT and EmmaX-SKAT is O(N^3), where M_1 is the number of markers used for estimating the full GRM and N is the sample size. In Step 2, the association test for each gene costs O(qK) in SAIGE-GENE, where q is the number of markers in the gene and K is the number of non-zero elements in the sparse GRM. Compared to O(qN^2) in Step 2 of SMMAT and EmmaX-SKAT, SAIGE-GENE decreases the computation time dramatically. For example, in the UK Biobank (N =408,910) with the relatedness coefficient ≥ 0.125 (corresponding to preserving samples with 3rd degree or closer relatives in the GRM), K = 493,536, which is the same order of magnitude of N, and hence O(qK) is greatly smaller than $O(qN^2)$. As the computation time in Step 2 is approximately linear to q, the number of markers in each variant set, the total computation time for exome-wide gene-based tests was projected by different q and plotted in **Figure S3.8**. In addition, we plotted the projected computation time for genome-wide region-based tests against the sample size as shown in **Figure S3.9**, in which 286,000 chunks with 50 markers per chunk were assumed to be tested, corresponding to 14.3 million markers in HRC-imputed UK Biobank data with MAF ≤ 1% and imputation info score ≥ 0.8.

With M_1 = 340,447, it takes SAIGE-GENE 2,238 CPU hours for exome-wide gene-based tests and 3,919 CPU hours for genome-wide region-based tests for waist hip ratio with N = 400,000 and each test contains 50 markers on average. Compared to EmmaX-SKAT and SMMAT, SAIGE-GENE is 25 times faster for exome-wide gene-based tests and 161 times faster for genome-wide region-based tests. More details about the computation cost are presented in **Table S3.1**.

To evaluate whether the additional steps in the robust adjustment for binary traits increases computation cost, we have obtained computation time of SAIGE-GENE with and without the

adjustment when analyzing the UK Biobank data for glaucoma (PheCode:365). Samples were randomly selected from 4,462 glaucoma cases and 397,701 controls respectively, so the case-control ratio remained the same in sub-sampled data sets. The results are presented in **Table S3.2** and plotted in **Figure S3.10,** showing that the robust adjustment only slightly increases the computation cost (1,269 vs 1,232 CPU hours for exome-wide analysis with M_1 = 93,511) compared to the unadjusted approach.

The computation time for constructing the sparse GRM is O($M_1^* N^2$ + $M_1 K$). M_1^* is the number of a small set of markers used for initial determination of related sample pairs based on a relationship coefficient cutoff, which by default is set to be 2,000. This step is only needed for each data set for one time to create a sparse GRM and the constructed sparse GRM will be re-used for all phenotypes in the same cohort or biobank. For example, for the UK Biobank with N = 408,910, M_1= 340,447, M_1^* = 2000, K = 493,536 with the relationship coefficient $\geq$ 0.125, corresponding to up to 3rd degree relatives, it took 312 CPU hours to create the sparse GRM. Parallel computation is allowed for this step.

3.3.2 Gene-based association analysis of quantitative traits in HUNT and UK Biobank

We applied SAIGE-GENE to analyze 13,416 genes, with at least two rare (MAF $\leq$ 1%) missense and stop-gain variants that were directly genotyped or imputed from HRC for high-density lipoprotein (HDL) in 69,716 Norwegian samples from a population-based Nord Trøndelag Health Study (HUNT)[9]. The HUNT study has substantial sample relatedness, in which ~55,000 samples have at least one up to 3rd degree relatives. The quantile-quantile (QQ) plot for the p-values of SKAT-O tests from SAIGE-GENE for HDL in HUNT is shown **Figure 3.2A.** As **Table 3.1** shows, eight genes reached the exome-wide significant threshold (p-value $\leq$ 2.5x10^{-6}) and all of them are

located in the previously reported GWAS loci for HDL(Willer et al., 2008; Willer et al., 2013). By extending 500kb up and down stream, a top significant hit from single-variant association tests has been identified around each gene. For genes *LIPC*, *LIPG*, *NR1H3*, and *CKAP5*, the top hits are common variants with MAF > 5% and the top hits in *FSD1L*, *ABCA1* and *RNF111* are less frequent non-coding variants that are not included in the gene-based tests. After conditioning on top hits, all genes, except for *FSD1L*, remained exome-wide significant, suggesting that SAIGE-GENE has identified associations of rare coding variants of those genes that are independent from the nearby association signals, pointing to candidate causal genes at those loci.

Table 3.1 Genes that are significantly associated with automated read pulse rate and glaucoma in the UK Biobank and high-density lipoprotein (HDL) in the HUNT study with SKAT-O p-values < $2.5x10^{-6}$ from SAIGE-GENE. Conditional analysis was performed when the top hit in the locus (+/- 500kb of the start and end positions of the gene) is not included in the gene-based test. The p-value of conditional analysis is NA when the top hit is a rare missense or stop gain variant included in the gene-based test.

	Gene	Number of Markers	SAIGE SKAT-O Test		Top Hit in the Locus		
			p-value	p-value Conditional	Variant (*GRCh37/hg19*)	p-value	MAF
Pulse Rate (UK Biobank)	*TBX5*	4	9.69E-35	NA	12:114837349_C:A	7.73E-35	0.0049
	MYH6	14	3.61E-15	2.56E-13	14:23861811_A:G	1.04E-168	0.3698
	TTN	368	3.18E-10	3.41E-06	2:179721046_G:A	8.73E-100	0.0885
	KIF1C	12	4.78E-10	NA	17:4925475_C:T	3.18E-10	0.0063
	ARHGEF40	7	7.02E-08	2.57E-10	14:21542766_A:G	3.30E-52	0.1688
	FNIP1	8	3.58E-07	4.31E-02	5:131107733_C:T	1.22E-08	0.0027
	DBH	12	1.74E-06	1.74E-06	9:136149399_G:A	3.46E-06	0.1870
HDL (HUNT)	*LCAT*	3	7.34E-50	NA	16:67974303_A:T	1.78E-48	0.0008
	LIPC	4	1.25E-29	6.63E-31	15:58723939_G:A	7.50E-89	0.1889
	FSD1L	3	7.40E-15	1	9:107793713_T:C	1.45E-20	0.0021
	ABCA1	14	3.32E-11	1.28E-11	9:107620797_A:G	3.64E-48	0.0055
	LIPG	3	2.15E-10	2.41E-10	18:47156926_C:A	5.92E-40	0.2348
	NR1H3	2	6.53E-09	1.69E-09	11:47246397_G:A	3.66E-13	0.322
	CKAP5	7	1.62E-08	1.21E-09	11:47246397_G:A	3.66E-13	0.322
	RNF111	11	1.18E-07	1.37E-09	15:58856899_C:G	2.82E-24	0.0047
Glaucoma (UK Biobank)	*MYOC*	6	1.23E-06	NA	1:171605478_G:A	9.13E-16	0.00137

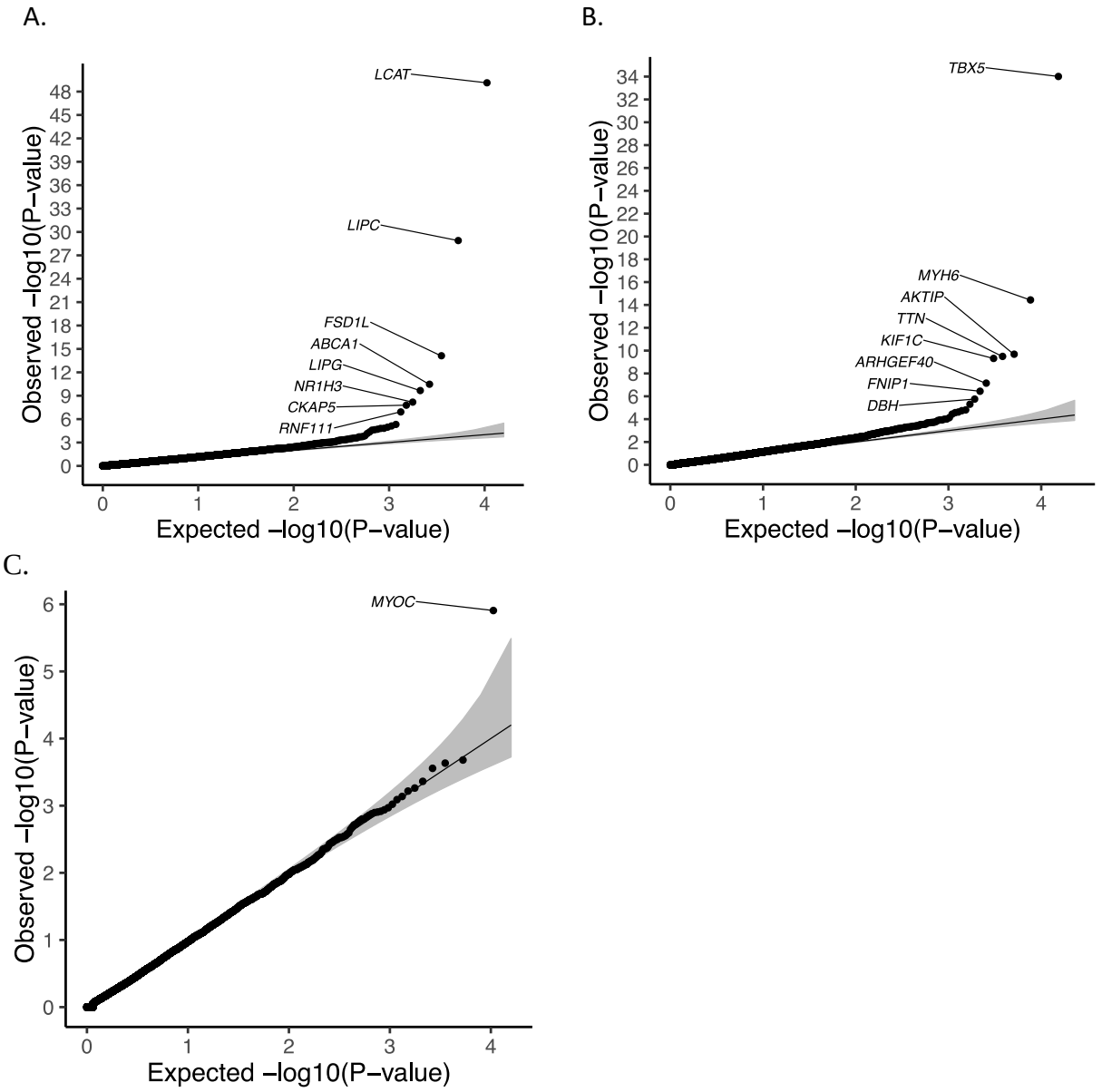


Figure 3.2 Quantile-quantile plots of exome-wide gene-based association results for A. high-density lipoprotein (HDL) in the HUNT study (N = 69,214). SKAT-O approach in SAIGE-GENE was performed for 13,416 genes with stop-gain and missense variants with MAF ≤ 1%, of which 10,600 having at least two variants are plotted. B. automated read pulse rate in the UK Biobank (N = 385,365). C. glaucoma in the UK Biobank (N cases = 4,462; N controls = 397,761). SKAT-O approach in SAIGE-GENE was performed for 15,338 genes with stop-gain and missense variants with MAF ≤ 1%, of which 12,638 having at least two variants are plotting.

We also applied SAIGE-GENE to analyze 15,342 genes for 53 quantitative traits using 408,910 UK Biobank participants with White British ancestry(C. Bycroft et al., 2018). Heritability estimates based on the full GRM are presented in **Table S3.3A**. **Table S3.4A** presents all genes with p-values reaching the exome-wide significant threshold ($p \leq 2.5x10^{-6}$). The same MAF cutoff $\leq 1\%$, for missense and stop-gain variants were applied**. Figure 3.2B** shows the QQ plot for automated read pulse rate as an exemplary quantitative phenotype in the UK Biobank. After conditioning on the most significant nearby variants, *MYH6*, *ARHGEF40* and *DBH* remain significant (**Table 1**). Gene *TBX5, MYH6, TTN,* and *ARHGEF40* are known genes for heart rates by previous GWAS studies(Arking et al., 2014; Eijgelsheim et al., 2010; Eppinga et al., 2016; Holm et al., 2010). To our knowledge, *KIF1C* and *DBH* have not been reported by association studies for heart rates, but both homozygous and heterozygous *DBH* mutant mice have decreased heart rates(Swoap et al., 2004). For the gene *DBH*, no single variant reaches the genome-wide significant threshold (the most significant variant is 9:136149399 (GRCh37) with MAF = 18.7% and p-value $=3.46x10^{-6}$).

In the analysis of all 53 quantitative traits in the UK Biobank, 199 gene-phenotype pairs were significant at exome-wide significant threshold ($p \leq 2.5x10^{-6}$). Among them fifteen genes for fourteen phenotypes were not significant by the single variant test, as the most significant single-variant association p-value in each of these loci (500kb up and down stream around each gene) did not reach the genome-wide threshold (p-value $< 5x10^{-8}$)(**Table S3.5**). For example, *TBX5*, which has been previously reported to be associated with heart rates(Holm et al., 2010), was significant by SAIGE-GENE for the automated read pulse rate ($p\text{-value}_{SKAT\text{-}O} = 2.87x10^{-7}$). However, the top variant in the locus was not genome-wide significant (p-value = $2.91x10^{-7}$). *ARID1B* has been previously reported to be associated with blood pressure in individuals with African ancestry(Sung

et al., 2018) and identified by SAIGE-GENE for automated read mean of diastolic blood pressure ($p\text{-value}_{SKAT\text{-}O} = 1.08x10^{-6}$), while the most significant single variant association p-value was $9.01x10^{-7}$. In addition, SAIGE-GENE has identified several potentially novel gene-phenotype associations, including *DBH* for automated read pulse rate ($p\text{-value}_{SKAT\text{-}O} = 1.74x10^{-6}$), *C10orf35* for body fat percentage ($p\text{-value}_{SKAT\text{-}O} = 3.64x10^{-7}$), a gene have been reported to be associated with type 2 diabetes(Replication et al., 2014) and blood lipids by previous GWAS(Bandesh et al., 2019). After conditioning on the most significant nearby variants, total 64 genes for 12 traits remained exome-wide significant (**Table S3.6A).** Our results have successfully replicated several previous findings, such as the association between the rare coding variants of *ADAMTS3* and height(Marouli et al., 2017), *ZFAT* and height(Marouli et al., 2017), and *RRAS* and blood pressure(Surendran et al., 2016). These results have demonstrated the value of gene-based tests for identifying genetic factors for complex traits.

3.3.3 Gene-based association analysis of binary traits in UK Biobank

We also applied SAIGE-GENE to ten binary phenotypes with various case-control ratios in the UK Biobank. The heritability estimates in a liability scale are presented in **Table S3.3B**. Nine genes for six binary phenotypes reached the exome-wide significant threshold ($p\text{-value} < 2.5x10^{-6}$) (**Table S3.4B**), all of which have been identified by both SAIGE-GENE and single variant tests, including the gene *MYOC*, known for glaucoma(Turalba & Chen, 2008) (**Figure 3.2C**). Six genes for six binary phenotypes remained exome-wide significant after conditioning on top variants (**Table S3.6B).** Gene *GORASP1,* encoding Golgi Reassembly Stacking Protein 1 involved in the vesicle-mediated transport pathway, remained significant after conditioning on the top hit for diseases of hair and hair follicles.

3.3.4 Simulation Studies

We investigated the empirical type I error rates and power of SAIGE-GENE through simulation. We followed the steps described in the Online Methods section to simulate genotypes and phenotypes for 10,000 samples in two settings. One has 500 families and 5,000 unrelated samples and the other one has 1,000 families, each with 10 family members based on the pedigree shown in **Figure S3.11.**

3.3.4.1 Type I error rates

The type I error rates of SAIGE-GENE, EmmaX-SKAT, and SMMAT have been evaluated based on gene-based association tests performed on 10^7 simulated gene-phenotype combinations, each with 20 genetic variants with MAF $\leq$ 1% on average. A sparse GRM with a cutoff 0.2 for the coefficient of relatedness was used in SAIGE-GENE. Two different values of variance component parameter corresponding to the heritability h^2 = 0.2 and 0.4 were considered for continuous traits, respectively. The empirical type I error rates at the α = 0.05, 10^{-4} and 2.5×10^{-6} are shown in the **Table S3.7**. Our simulation results suggest that SAIGE-GENE has relatively well controlled type I error rates, while the type I error rates are slightly inflated when heritability is relatively high (h^2 = 0.4). Similar results have been observed on a larger sample size with 1,000 families and 10,000 unrelated samples (**Supplementary Materials 3.5.2.1** and **Table S3.8**). Adjusting the test statistics using the genomic control (GC) inflation factor lambda has addressed the inflation (**Supplementary Materials 3.5.1.7)**.

Further simulations have been conducted to evaluate type I error rates of SAIGE-GENE, EmmaX-SKAT, and SMMAT for skewed distributed phenotypes, which are common in real data (**Figure S3.12A**). All three methods had inflated type I error rates for phenotypes having skewed distribution (**Table S3.9**). With inverse normal transformation on phenotypes (**Figure S3.12B**), the inflation has been dramatically reduced but slight inflation was still observed (**Table S3.9**). A

potential reason is that inverse normal transformation disrupts sample relatedness in raw phenotypes, leading to poor fitting for the null GLMM. We then conducted a three-step phenotype transformation procedure as described in **Supplementary Materials 3.5.2.2**, which maintains sample relatedness in raw phenotypes, and all three methods then have well controlled type I error rates (**Table S3.10**). From simulation studies using real genotype data from the UK Biobank, we show that SAIGE-GENE well controlled type I error rates in the presence of subtle population structure or non-negligible cryptic relatedness between families (**Table S3.11 and S3.12**). Details have been described in **Supplementary Materials 3.5.2.3** and **3.5.2.4**.

We have also evaluated the empirical type I error rates of SAIGE-GENE for binary traits with various case-control ratios. Similar with continuous traits, a sparse GRM with a cutoff 0.2 for the coefficient of relatedness was used. The variance component parameter $\tau = 1$ was assumed, corresponding to liability-scale heritability 0.23. As expected, when case-control ratios were balanced or moderately unbalanced (e.g. 1:1 and 1:9), type I error rates were well controlled even without the robust adjustment, while when the ratios were extremely unbalanced (e.g. 1:19 and 1:99), inflation was observed (**Table S3.13A and Figure S3.6**). With the robust adjustment combining SPA and ER, type I error rates were relatively well controlled in the presence of unbalanced case-control ratios (**Table S3.13B and Figure S3.6**). However, for phenotypes with case-control ratio=1:99, slight inflation was still observed, although the inflation has been dramatically alleviated compared to the unadjusted method. Then the genomic control adjustment can be used to further control the type I error rates (**Table S3.13B**). We have also evaluated empirical type I error rates of SAIGE-GENE for binary traits under case-control sampling with case-control ratios 1:1 and 1:9 based on a disease prevalence 1% in the population (**Supplementary Materials 3.5.2.5**) and observed well-controlled type I error rates (**Table S3.14**).

3.3.4.2 Power

Next, we evaluated empirical power of SAIGE-GENE and EmmaX-SKAT for quantitative traits. Two different settings of proportions of causal variants were used: 10% and 40%. In each setting, among causal variants, 80% and 100% have negative effect sizes. The absolute effect sizes for causal variants are set to be $|0.3\log_{10}(MAF)|$ and $|\log_{10}(MAF)|$, respectively, when the proportions of causal variants are 0.4 and 0.1. **Table S3.15** shows that the power of both methods is nearly identical for all simulation settings for Burden, SKAT and SKAT-O tests.

We have also evaluated empirical power of SAIGE-GENE for binary traits using two different study designs: cohort study with various disease prevalence (0.01-0.5); and case-control sampling with different case-control ratios (1:1-1:19) based on a disease prevalence 1% in the population. In each setting, 40% variants are simulated as causal variants. Among them, 80% are risk-increasing variants and 20% are risk-decreasing. The absolute effect sizes of causal variants are set to be $|0.55\log_{10}(MAF)|$ and $|0.35\log_{10}(MAF)|$ for cohort study and case-control sampling, respectively. **Table S3.16** shows the empirical power of SKAT-O in both simulation studies. SAIGE-GENE had similar empirical power as unadjusted SAIGE-GENE in balanced case-control ratios and higher power in unbalanced scenarios. The power is small when case: control ratio is 1:99 due to the limited number of cases (100 cases), which can be alleviated with larger sample size.

3.4 Discussion

In summary, we have presented a method, SAIGE-GENE, to perform gene- or region-based association tests in large cohorts or biobanks in the presence of sample relatedness. Similar to SAIGE(W. Zhou et al., 2018), which was previously developed by our group for single-variant

association tests, SAIGE-GENE uses generalized linear mixed models to account for sample relatedness, scalable computational approaches for large sample sizes, and the robust adjustment(Zhao et al., 2019) to account for unbalanced case-control ratios of binary traits .

SAIGE-GENE uses several optimization strategies that are similar to those used in SAIGE to make fitting the null GLMM feasible for large sample sizes. For example, instead of storing the genetic relationship matrix (GRM) in the memory, SAIGE-GENE stores genotypes that are used for constructing the matrix in a binary vector and computes the elements of the matrix as needed. Preconditioned conjugate gradient algorithm is also used to solve linear systems instead of the Cholesky decomposition. However, some optimization approaches are specifically applied in the gene-based tests in regard of rare variants. As estimating the variances of score statistics for rare variants are more sensible to family structures, we use a sparse GRM to preserve close family structures rather than ignoring all sample relatedness. In addition, the variance ratios are estimated for different minor allele count (MAC) categories, especially for those extremely rare variants with MAC lower than or equal to 20.

For binary phenotypes, SAIGE-GENE applies the robust adjustment combining SPA and ER, thereby also relatively well controls the type I error rates for both balanced and unbalanced case-control phenotypes. However, slight inflation is still observed in extremely unbalanced phenotypes ($\leq$1:99). To address this possible issue, we suggest using the genomic control to further control type I error.

In numerical optimization, using good initial values can improve the model convergence. In the analysis of 24 quantitative traits in the UK Biobank with sample size (N) $\geq$ 100,000, we note that the models with the full GRM and the sparse GRM produced different variance component estimates, but they are relatively concordant (Pearson's correlation R^2 = 0.66, **Figure S3.13**). This

indicates that the parameter estimates from the sparse GRM can be used as initial values to facilitate the model fitting. We implemented this approach in SAIGE-GENE.

SAIGE-GENE has some limitations. First, similar to SAIGE and other mixed-model methods, the time for algorithm convergence to fit the generalize linear mixed models may vary among phenotypes and study samples given different heritability levels and sample relatedness. Second, similar to SAIGE(W. Zhou et al., 2018) and SMMAT(H. Chen et al., 2019), SAIGE-GENE uses penalized quasi-likelihood (PQL)(Breslow & Clayton, 1993) for binary traits to estimate the variance component in binary phenotypes which is known to be biased. However, as shown in simulation studies in SAIGE(W. Zhou et al., 2018) and SMMAT(H. Chen et al., 2019), PQL-based approaches works well to adjust for sample relatedness.

Overall, we have shown that SAIGE-GENE can account for sample relatedness while maintaining test power through extensive simulation studies. By applying SAIGE-GENE to the HUNT study[9] and the UK Biobank(C. Bycroft et al., 2018) followed by conditioning on most significant variants in the testing loci, we have demonstrated that SAIGE-GENE can identify potentially novel association signals that are independent from the nearby association signals from the single-variant tests. Currently, our method is the only available mixed effect model approach for gene- or region-based rare variant tests for large sample data, while accounting for unbalanced case-control ratios for binary traits. By providing a scalable solution to the current largest and future even larger datasets, our method will contribute to identifying trait-susceptibility rare variants and genetic architecture of complex traits.

URLs

SAIGE (version 0.35.8.8), https://github.com/weizhouUMICH/SAIGE/.

SMMAT (version 1.0.2), https://github.com/hanchenphd/GMMAT.

EmmaX-SKAT (SKAT version_1.3.2.1), https://cran.r-project.org/web/packages/SKAT/index.html.

UK-Biobank analysis results (Gene-based summary statistics for 53 quantitative phenotypes in the UK Biobank by SAIGE-GENE), https://www.leelabsg.org/resources.

Code and data availability

SAIGE-GENE is implemented as an open-source R package available at https://github.com/weizhouUMICH/SAIGE/master.

The summary statistics and QQ plots for 53 quantitative phenotypes and 10 binary phenotypes in UK Biobank by SAIGE-GENE are currently available for public download at https://www.leelabsg.org/resources.

3.5 Supplementary Materials

3.5.1 Algorithm details

3.5.1.1 Step 0. Constructing the sparse GRM

In the sparse GRM, denoted by ψ_s, GRM elements below a user-specified relative coefficient cutoff are zeroed out with close family structures preserved. To improve the test accuracy for rare variants, SAIGE-GENE approximates the variance of score statistics calculated with the full GRM ψ_f using the variance calculated with the sparse GRM ψ_s and the ratios of these two variance estimates estimated using a subset of genetic markers.

To construct the sparse GRM ψ_s, a small subset of randomly select markers were used to identify related sample pairs whose relative coefficient pass the use-specified cutoff, which is to find out the indices of non-zero elements in ψ_s. Next, the values of the nonzero elements in ψ_s are then estimated using the full set of genetic markers that are used in Step 1 for ψ_f. This step is only needed for once for each data set or biobank and parallel computation is allowed. Once the sparse GRM is constructed for a data set, it can be re-used in SAIGE-GENE for all phenotypes.

3.5.1.2 Step 1. Fitting the null generalized linear mixed model

The same model fitting framework and computation approaches used in the original SAIGE(W. Zhou et al., 2018) are used in SAIGE-GENE to fit the null GLMM for large sample sizes. These include estimating model parameters using the AI-REML approach(Gilmour et al., 1995; Lee & van der Werf, 2006), solving linear systems by the preconditioned conjugate gradient method(Kaasschieter, 1988), using Hutchinson's randomized trace estimator(Avron & Toledo, 2011; Hutchinson, 1990) to obtain traces of matrices, and allowing for parallel computation for

the vector multiplication. For AI-REML, we particularly used the approach in GMMAT(H. Chen et al., 2016; Gilmour et al., 1995) to calculate the average information without performing the *n* by *n* matrix inversion by using PCG. For details of the likelihood, parameter estimates and information matrices, please refer to the Supplementary Note in the SAIGE paper(W. Zhou et al., 2018). In addition, SAIGE-GENE can estimate variance component parameters, thus heritability, by fitting a null GLMM using the sparse GRM. The estimated variance component parameters can then be used as initial values for the model fitting with the full GRM, which can be a better approach than using a randomly chosen initial value. By plotting the heritability estimates using the sparse GRM versus using the full GRM for 24 quantitative traits with sample size larger than or equal to 10,000 from the UK Biobank (**Figure S3.13**), we have shown that variance component estimates from the full and sparse GRMs are relatively concordant (Pearson's correlation R^2 = 0.66). For real-data analysis, robust performance of convergence has also been observed for some phenotypes, such as waist hip ratio (N = 408,144) in the UK Biobank. Using initial values 0.5 for heritability, step 1 did not even converge after 6,300 CPU hours, while using initial values estimated with the sparse GRM, it took 1836 CPU hours to finish the step 1.

3.5.1.3 Step 2. Gene-based association tests

Test statistics of the Burden, SKAT and SKAT-O tests for a gene can be constructed based on the score statistics from the marginal model for individual variants in the gene. Suppose there are q variants in the region or gene to test. The score test statistics for variant j (j=1,..,q) under H_o: $\beta_j = 0$ is $T_j = {g_j}^T(Y - \hat{\mu})$ where g_j and Y are $N \times 1$ genotype and phenotype vectors, respectively, and $\hat{\mu}$ is the estimated mean of Y under the null hypothesis.

Let u_j denote a threshold indicator or weight for variant j and U be a diagonal matrix with u_j as the jth element. The Burden test statistics can be written as $Q_{Burden} = \left(\sum_{j=1}^{q} u_j T_j\right)^2$. Suppose $\tilde{G} = G - X(X^T\widehat{W}X)^{-1}X^T\widehat{W}G$, where $G = (g_1, \dots, g_q)$ is the $N \times q$ genotype matrix of the q genetic variants, and $\hat{P} = \hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}X\left(X^T\hat{\Sigma}^{-1}X\right)^{-1}X^T\hat{\Sigma}^{-1}$ with $\hat{\Sigma} = \widehat{W}^{-1} + \hat{\tau}\psi$. Under the null hypothesis of no genetic effects, Q_{Burden} followed $\lambda_B\chi_1^2$, where $\lambda_B = J^T\mathrm{U}\tilde{G}^T\hat{P}\tilde{G}\mathrm{U}J$ and J is a $q \times 1$ vector with all elements being unity and $\chi_1^2$is a chi-squared distribution with 1 degree of freedom(S. Lee et al., 2014). The SKAT test(M. C. Wu et al., 2011) can be written as $Q_{SKAT} = \sum_{j=1}^{q} u_j^2 T_j^2$, which follows a mixture of chi-square distribution $\sum_{j=1}^{q} \lambda_{Sj}\chi_1^2$, where λ_{Sj} are the eigenvalues of $U\tilde{G}^T\hat{P}\tilde{G}U$. The SKAT-O test developed by Lee et al in 2012 (S. Lee et al., 2012) uses a linear combination of the Burden and SKAT tests statistics $Q_{SKATO} = (1-\rho)Q_{SKAT} + \rho Q_{Burden}, 0 \leq \rho \leq 1$. To conduct the test, the minimum p-value from grid of ρ is calculated and the p-value of the minimum p-value is estimated through numerical integration. Following the suggestion in Lee *et al*(Lee et al., 2013), we use a grid of eight values of $\rho = (0, 0.1^2, 0.2^2, 0.3^2, 0.4^2, 0.5^2, 0.5, 1)$ to find the minimum p-value.

3.5.1.4 Estimating $\widetilde{\boldsymbol{G}}^{\boldsymbol{T}}\widehat{\boldsymbol{P}}\widetilde{\boldsymbol{G}}$

For each gene, given $\hat{P}$, calculation of $\tilde{G}^T\hat{P}\tilde{G}$ can be computationally expensive Suppose $\tilde{g} = g - X\left(X^T\widehat{W}X\right)^{-1}X^T\widehat{W}g$, which represents a covariate adjusted single variant genotype $N \times 1$ vector vector. To reduce computation cost, an approximation approach has been used in SAIGE(W. Zhou et al., 2018), BOLT-LMM(Loh et al., 2015) and GRAMMAR-GAMMAR(Svishcheva et al., 2012), in which the ratio between $\tilde{g}^T\hat{P}\tilde{g}$ and $\tilde{g}^T\tilde{g}$ is estimated by a small subset of randomly selected genetic markers that has been shown to be approximately constant for all variants(W. Zhou et al.,

2018). Given the ratio $\hat{r} = \tilde{g}^T \hat{P} \tilde{g} \,/ \tilde{g}^T \tilde{g}$, $\tilde{g}^T \hat{P} \tilde{g}$ for all other variants can be easily obtained as $\hat{r} \tilde{g}^T \tilde{g}$. However, the variations of estimated $\hat{r}$ for extremely rare variants are large and including some closely related samples in the denominator helps reduce the variation of $\hat{r}$ as shown in **Figure S3.2**. It can also be observed from the plots in **Figure S3.2** that the variance ratio for those extremely rare variants could be quite different from the ratio for more frequent variants, so SAIGE-GENE estimates variance ratios for different MAC categories. By default, MAC categories are set to be MAC equals to 1, 2, 3, 4, 5, 6 to 10, 11 to 20, and is greater than 20. For each MAC category, a ratio $\hat{r}_s$ is estimated as the average of the ratios computed from 30 randomly selected markers, among which every marker has a ratio $\tilde{g}^T \hat{P} \tilde{g} \,/ \tilde{g}^T \hat{P}_s \tilde{g}$, where $\hat{P}_s = \hat{\Sigma}_s^{-1} - \hat{\Sigma}_s^{-1} X \left(X^T \hat{\Sigma}_s^{-1} X \right)^{-1} X^T \hat{\Sigma}_s^{-1}$ and $\hat{\Sigma}_s = \widehat{W}^{-1} + \tau \psi_s$. ψ_S is a sparse GRM that preserves closely related samples. The coefficient of variance (CV) of $\hat{r}$ is used to evaluate the numerical stability of the $\hat{r}$ estimation. As in SAIGE, the default value of CV threshold is 0.001. If CV of $\hat{r}$ is larger than the threshold, SAIGE-GENE will increase the number of markers by 10 to estimate $\hat{r}$ until the estimation is stable with CV below or equal to the threshold. Once the variance ratios have been estimated for different MAC categories. For each genetic marker in genes or regions that are to be tested in Step 2, a $\hat{r}_s$ can be obtained according to its MAC. Let $\hat{R}_s$ be a $q \times 1$ vector whose jth element is the ratio $\hat{r}_s$ for the jth marker in the tested gene. For the tested gene with q markers, $\tilde{G}^T \hat{P} \tilde{G}$ can be approximated as $\hat{R}_s^{\frac{1}{2}} \tilde{G}^T \hat{P}_s \tilde{G} \hat{R}_s^{\frac{1}{2}}$.

Products of $\hat{\Sigma}_s^{-1}$and other vectors or matrices are obtained using the sparse LU decomposition through the solve function in R(Davis, 2006). Note that the $N \times N$ matrix $\hat{P}_s$ is not a sparse matrix

because $\hat{\Sigma}_s^{-1}$ is not sparse and $\tilde{G}$ is also a dense matrix, which is converted from the sparse matrix G, as $\tilde{G} = G - X\left(X^T\widehat{W}X\right)^{-1}X^T\widehat{W}G$. It can be shown that

$$\tilde{G}^T\hat{P}_s\tilde{G} = (G^T - G^T\widehat{W}X\left(X^T\widehat{W}X\right)^{-1}X^T)\left(\hat{\Sigma}_s^{-1} - \hat{\Sigma}_s^{-1}X\left(X^T\hat{\Sigma}_s^{-1}X\right)^{-1}X^T\hat{\Sigma}_s^{-1}\right)(G$$

$$- X\left(X^T\widehat{W}X\right)^{-1}X^T\widehat{W}G) = G^T\hat{P}_sG$$

Computation of $G^T\hat{P}_sG$ is more computational efficient than that of $\tilde{G}^T\hat{P}_s\tilde{G}$ as G can be stored as a sparse matrix.

3.5.1.5 Conditional analysis

To test whether the association signals from a tested gene or region are independent from a given marker or multiple markers, the conditional analysis based on summary statistics from unconditional association tests with the linkage disequilibrium r^2 among testing and conditioning markers(Liu et al., 2014) have been implemented in SAIGE-GENE.

Let G be the genotypes for a gene to be tested for association, which contains q markers, and G_2 be the genotypes for the conditioning markers, which contains q_2 markers. Let β denote a $q \times 1$ coefficient vector of the genetic effect for the gene to be tested and β_2 be a $q_2 \times 1$ coefficient vector of the genetic effect for the conditioning markers. The genotype matrix with the non-genetic covariates projected out $\tilde{G} = G - X\left(X^T\widehat{W}X\right)^{-1}X^T\widehat{W}G$ and $\tilde{G}_2 = G_2 - X\left(X^T\widehat{W}X\right)^{-1}X^T\widehat{W}G_2$. In the unconditioned association tests, the test statistics $T = \tilde{G}^T(Y - \hat{\mu})$ and $T_2 = \tilde{G}_2^{\ T}(Y - \hat{\mu})$. In conditional analysis, under the null hypothesis, E(T) = E($\tilde{G}^T P(\tilde{G}_2\beta_2)) = \tilde{G}^T\hat{P}\tilde{G}_2\beta_2$ and E(T_2) =

E($\tilde{G}_2^T P(\tilde{G}_2\beta_2)) = \tilde{G}_2^T \hat{P}_s \tilde{G}_2 \beta_2$. T and T_2 jointly follow the multivariate normal with mean (E(T), E(T_2)) and variance $S = \begin{bmatrix} \tilde{G}^T \hat{P} \tilde{G} & \tilde{G}^T \hat{P} \tilde{G}_2 \\ \tilde{G}_2^T \hat{P} \tilde{G} & \tilde{G}_2^T \hat{P} \tilde{G}_2 \end{bmatrix}$.

Thus under the null hypothesis of β=0, the $T|T_2$ follows the conditional distribution E($T|T_2$) = $\tilde{G}^T \hat{P} \tilde{G}_2 \, (\tilde{G}_2^T \hat{P} \tilde{G}_2)^{-1} T_2$ and var($T|T_2$) = $\tilde{G}^T \hat{P} \tilde{G} - \tilde{G}^T \hat{P} \tilde{G}_2 \left(\tilde{G}_2^T \hat{P} \tilde{G}_2\right)^{-1} \tilde{G}_2^T \hat{P} \tilde{G}$ = $G^T \hat{P} G - G^T \hat{P} G_2 (G_2^T \hat{P} G_2)^{-1} G_2^T \hat{P} G$. The test statistic of the conditional analysis can be written as (T - E($T|T_2$))^2/ var($T|T_2$), which follows the χ^2distribtuion with one degree of freedom. Similar to the unconditional analysis, we approximate these terms by $\tilde{G}^T \hat{P}_s \tilde{G}, \tilde{G}^T \hat{P}_s \tilde{G}_2, \tilde{G}_2^T \hat{P}_s \tilde{G}_2$, and $\tilde{G}_2^T \hat{P}_s \tilde{G}$ and the corresponding variance ratio matrices.

3.5.1.6 Robust adjustment to account for unbalanced case-control ratios in binary traits

To account for unbalanced case-control ratios in binary traits, we use a recently developed robust adjustment approach with a simple modification(Zhao et al., 2019). Note that the robust adjustment was developed for independent samples. It uses saddlepoint approximation (SPA)(Daniels, 1954; R. Dey et al., 2017; Kuonen, 1999) and efficient resampling (ER)(Lee et al., 2016) to obtain accurate single variant association P-value for the jth variant and then calibrates the variance of score statistics T_j. SPA is a statistical method to calculate the distribution function using the cumulant generating function (CGF). Suppose $K_j(t)$ is the CGF of the score statistic T_j, which can be derived based on the fact that $Y_i \sim \text{Bernoulli}(\mu_i)$ under the null. For independent samples, the estimation of $K_j(t)$ is

$$\hat{K}_j(t; \hat{\mu}, c) = \sum_{i=1}^{N} \log\left(1 - \hat{\mu}_i + \hat{\mu}_i e^{t\widetilde{G_i}}\right) - t \sum_{i=1}^{N} \widetilde{G}_i \hat{\mu}_i,$$

where $\widehat{\mu_i}$ is the estimation of μ_i from the null model, and $\widetilde{G_i}$ is the covariate adjusted genotype vector. To account for sample relatedness, we use the original SAIGE, which adapts the SPA to GLMM. Under the GLMM, we use the following CGF

$$\widehat{K_J}(t;\hat{\mu},c) = \sum_{i=1}^{N} \log\left(1-\widehat{\mu_i}+\widehat{\mu_i}e^{ct\widetilde{G_i}}\right) - ct\sum_{i=1}^{N}\widetilde{G_i}\widehat{\mu_i},$$

where $c = \mathrm{Var}^*(T_j)^{-1/2}$ and $\mathrm{Var}^*(T_j) = \widetilde{G_j}^T W \widetilde{G_j}$ is a variance estimator without accounting the fact that the random effect b is estimated from data. Then, the distribution function of the score statistic T_j can be approximated by

$$\Pr(T_j < q) = \tilde{F}(q) = \Phi\left\{w + \frac{1}{w} log\left(\frac{v}{w}\right)\right\},$$

where $w = sgn(\hat{t})\sqrt{2\left(\hat{t}s - K_j(\hat{t})\right)}$, $v = \hat{t}\sqrt{K_j''(\hat{t})}$, $\hat{t}$ is the solution to the equation $K_j'(\hat{t}) = q$, and Φ is the distribution function of the standard normal distribution. For details, please refer to the SAIGE paper(W. Zhou et al., 2018).

Since SPA is an asymptotic based approach, it can provide incorrect p-values when MAC is very low (ex. MAC < 10). To address this issue, we use ER(Lee et al., 2016) for variants with MAC ≤ 10. ER is a resampling method that resamples the case–control status of individuals with a minor allele at a given variant given the disease risk μ_i. ER was developed under the assumption that samples are independent. Here we apply ER with the disease risk $\widehat{\mu_i}$ estimated by GLMM, using the fact that given random effects b_i, samples are independent. The dependency among samples are incorporated through the random effect estimates.

The remaining part is nearly identical as the robust method in independent samples(Zhao et al., 2019). Let $\hat{V} = diag(\hat{v}_1, \dots, \hat{v}_q)^T$ is a qxq diagonal matrix of the diagonal element of $\hat{R}_S^{\frac{1}{2}}\tilde{G}^T\hat{P}_s\tilde{G}\hat{R}_S^{\frac{1}{2}}$. We note that j-th diagonal element of $\hat{V}$ is the estimated variance of T_j. For each variant j, when the score statistic T_j lies outside of two standard deviations of the mean (i.e. zero), we apply SPA (when MAC > 10) or ER (when MAC ≤ 10) to calculate the p-value $\tilde{p}_j$, and calculate $\tilde{v}_j = T_j^2/\chi^2_{quantile}(1-\tilde{p}_j)$, where $\chi^2_{quantile}$ is the quantile function of the chi-square distribution with one degree of freedom. If T_j lies within the two standard deviations of the mean, we use $\tilde{v}_j = \hat{v}_j$. Suppose $\tilde{R} = diag(\tilde{v}_1/\hat{v}_1, \tilde{v}_2/\hat{v}_2, \dots, \tilde{v}_q/\hat{v}_q)$ is a qxq diagonal matrix. Now we use $\tilde{R}^{\frac{1}{2}}\hat{R}_S^{\frac{1}{2}}\tilde{G}^T\hat{P}_s\tilde{G}\hat{R}_S^{\frac{1}{2}}\tilde{R}^{\frac{1}{2}}$ as the estimate of $\tilde{G}^T\hat{P}\tilde{G}$ to calculate p-values and carry out conditional analysis.

3.5.1.7 Genomic Control (GC) for the further adjustment of p-values

To further control for type I error rates, SAIGE-GENE allows for using the genome control inflation factor. Let λ_{GC} be the genome control inflation factor from the gene-based test, which is obtained by converting p-values of gene-based test to χ^2_1 statistics. To better capture the inflation in tail areas we obtain λ_{GC} at p-value=0.05. And then we divide the λ_{GC} in the χ^2_1 statistics and then obtain the p-values using χ^2_1 distribution. As shown in **Table S3.7** and **Table S3.13**, this simple approach has successfully attenuated the type I error inflation.

3.5.1.8 Novel features in SAIGE-GENE compared to SMMAT

Same as SMMAT, SAIGE-GENE uses the logistic mixed model to conduct region- or gene-based association tests (Burden, SKAT and SKAT-O). Compared to SMMAT, SAIGE-GENE mainly has two improvements, which make it the only method so far that is feasible for large samples sizes, while accounting for case-control imbalance for binary phenotypes. It utilizes optimization

strategies as used in the original SAIGE for the scalability, including storing raw genotypes in a binary vector and elements of GRM are calculated when needed rather than being stored to reduce the memory usage, replacing the cholesky decomposition by the preconditioning conjugate gradient (PCG) to solve linear systems without calculating and inverting the $N \times N$ GRM, and approximating the variance of score statistics with the full GRM using the variance with a sparse GRM and the ratio of the two variances. To account for unbalanced case-control ratios for binary phenotypes, SAIGE-GENE uses a robust adjustment approach combing SPA(Daniels, 1954; R. Dey et al., 2017; Kuonen, 1999) and ER(Lee et al., 2016) as described in 1.3.3.

3.5.2 Additional simulation and real-data analysis results

3.5.2.1 Simulation studies with a larger sample size

Given that SAIGE-GENE has relatively well controlled type I error rates based on simulation studies (**Table S3.7**), it is noted that type I error rates are slightly higher for SAIGE-GENE than the other two methods. One potential reason for this could be that SAIGE-GENE uses several approximation approaches, to achieve feasibility for large data sets and a tradeoff exists between accuracy and computational efficiency for these approaches. To evaluate type I error rates with larger sample sizes, we conducted a simulation study containing 1,000 families and 10,000 independent samples, which doubles the sample sizes as in the original simulation with 500 families and 5,000 independent samples. With the heritability $h^2 = 0.2$, the empricial type I error rates for the Burden test, SKAT, and SKAT-O at three different α were estimated based on 10^7 tests, for which 1,000 randomly simulated phenotype sets and each was tested on 10,000 variant sets (**Table S3.8**). The type I error rates with the larger sample size (**Table S3.8**) are similar to those with a smaller sample size with 500 families and 5,000 independent samples (**Table S3.7**).

3.5.2.2 Simulation studies with skewed distributed phenotypes

In simulation studies, the phenotype y_i was generated to follow a normal distribution, but it can be skewed in real data. Hence, we have conducted additional simulation studies to evaluate the type I error rates of SAIGE-GENE in presence of skewed phenotypic distributions and compared to the other two methods. As described in the Data Simulation subsection of ONLINE METHODS**,** phenotypes were simulated from the following linear mixed model $y_i = X_1 + X_2 + G_i\beta + b_i + \varepsilon_i$, where G_i is the genotype value, β is the genetic effect sizes. Two covariates, X_1 and X_2, were simulated from Bernoulli(0.5) and N(0,1), respectively. b_i is the random effect simulated from $N(0, \tau\psi)$ and ε_i is the error term simulated from chi-square distribution with degree of freedom 1 and thereby the distribution of y is skewed (**Figure S3.12A**). 500 families and 5,000 independent samples were simulated and the empirical type I error rates for Burden test, SKAT, and SKAT-O at the three different α were estimated based on 10^7 tests, for which 1,000 randomly simulated phenotype sets and each was tested on 10,000 variant sets (**Table S3.9**). All three methods have inflated type I error rates for phenotypes having skewed distribution, especially in SKAT and SKAT-O tests. After the phenotypes were inverse normal transformed (**Figure S3.12B**), type I error rate inflation has been substantially reduced (**Table S3.9**).

Note that there is still slight inflation in all three methods after the inverse normal transformation on phenotypes, which can be because the inverse normal transformation may disrupt sample relatedness in the original phenotypes and thus impact the null model fitting. We then conducted a three-step phenotype transformation procedure, which is presumably able to avoid the issue above. Firstly, we fitted the null mixed model using raw skewed distributed phenotypes. Next, we conducted the inverse normal transformation on the residuals from step 1. Finally, transformed residuals were then used to fit another null mixed model, followed by gene- or region-based

association tests. As expected, using this three-step phenotype transformation procedure, the type I error rates are well controlled in both SAIGE-GENE and SMMAT for all Burden, SKAT, and SKAT-O tests (**Table S3.10**).

3.5.2.3 Simulation studies in the presence of population structure

To evaluate whether SAIGE-GENE can control type I error rates in the presence of subtle population stratification, we randomly selected 5,000 white British-UK samples and 5,000 non-UK European samples from the UK-Biobank after removing up to 3rd relatives. The phenotypes were simulated based on real genotypes of randomly selected L= 30,000 LD-pruned (r2 < 0.2) markers with MAF ≥ 1%. In particular, phenotypes were simulated following the model $y_i = X_{i1} + X_{i2} + \sum_{j=1}^{L} \hat{G}_{ij}\,\beta + \varepsilon_i$, where $\hat{G}_{ij}$ is the standardized genotype value for thejth marker of ith individual, β is the genetic effect size following $N(0, \tau/L)$, where $\tau = 0.2$, and ε_i is the error term simulated from $N(0, (1-\tau)I)$. Two covariates, X_{i1} and X_{i2}, were simulated from Bernoulli(0.5) and N(0,1). We conducted ~10,000 gene-based tests for each simulated phenotype set, replicated the simulation for 1,000 times. Note that we include the first 4 principal components, which were estimated for all European participants in the UK Biobank, as well as X_1 and X_2 as covariates in the linear mixed model. We evaluated the empirical type I error rates at the $\alpha = 0.05$, 10^{-4} and 2.5×10^{-6} as shown in **Table S3.11**, which suggest that SAIGE can produce well calibrated p-values in the presence of subtle population stratification.

3.5.2.4 Simulation studies in the presence of non-negligible cryptic relatedness between families

To evaluate whether SAIGE-GENE can control type I error rates in the presence of non-negligible cryptic relatedness between families, we have randomly selected 10,000 samples with white

British ancestry from UK Biobank. In particular, 5,000 samples were selected among those who are up to 3rd degree relatives and 5,000 sample were selected from the rest of the unrelated pool. Phenotypes were simulated using the same approach in Section 2.3. We conducted ~10,000 gene-based tests for each simulated phenotype set, replicated the simulation for 1,000 times. For each phenotype set, a null linear mixed model was fitted in Step 1 with covariates including the first 4 principal components, which were estimated for all White-British participants in the UK Biobank, and X_1 and X_2. We evaluated the empirical type I error rates at the $\alpha = 0.05$, 10^{-4} and $2.5x10^{-6}$ as shown in **Table S3.12**. These results have indicated that SAIGE can produce well calibrated type I error rates in the presence of non-negligible cryptic relatedness between families.

3.5.2.5 Simulation studies under case-control sampling

We have conducted additional simulation studies to evaluate the performance of SAIGE-GENE under case-control sampling. We simulated genotypes and phenotypes with prevalence 1% for 250,000 independent samples and 250,000 families, each with 10 family members (**Figure S3.11**) as an underlying large cohort. We then randomly selected 5,000 controls, together with the 5,000 cases, for a phenotype with case-control ratio 1:1. In addition, for a phenotype with case-control ratio 1:9, we randomly selected 500 cases and 9,500 controls from the large cohort. Empirical type I error rates of SAIGE-GENE have been evaluated based on the 10 million tests (**Table S3.14**), indicating that the type I error rates of SAIGE-GENE were well controlled under the case-control sampling. We have also evaluated the empirical power of SAIGE-GENE under case-control sampling through simulation studies as described in the RESULTS section of main texts. Results are presented in **Table S3.16,** showing that under case-control sampling, the empirical power of SAIGE-GENE with and without robust adjustment is similar when the case-control ratio is relatively balanced (1:1). As expected, SAIGE-GENE with robust adjustment has higher power in

unbalanced scenarios ($\leq 1:9$) than SAIGE-GENE without robust adjustment. Similar patterns have been observed for empirical power of SAIGE-GENE in simulations of a cohort study (**Table S3.16**). Note that the power in cohort study and case-control sampling is not directly comparable due to the difference effect sizes.

3.5.2.6 Exome-wide gene-based tests for automated read pulse rate in UK Biobank with different relatedness cutoffs in the sparse GRM

As a sensitivity analysis, we used another sample relatedness cutoff 0.2 for the sparse GRM to analyze automated read pulse rates in UK Biobank, which means all elements in the full GRM below 0.2 are zero'd in the sparse GRM. Scatter plots comparing p-values of the 15,342 genes with two different sample relatedness cutoff (0.125 and 0.2) are presented in **Figure S3.5**, showing highly concordant association p-values for all three gene-based tests: Burden, SKAT, and SKAT-O tests.

3.5.3 Supplementary figures

Figure S3.1. Workflow of SAIGE-GENE.

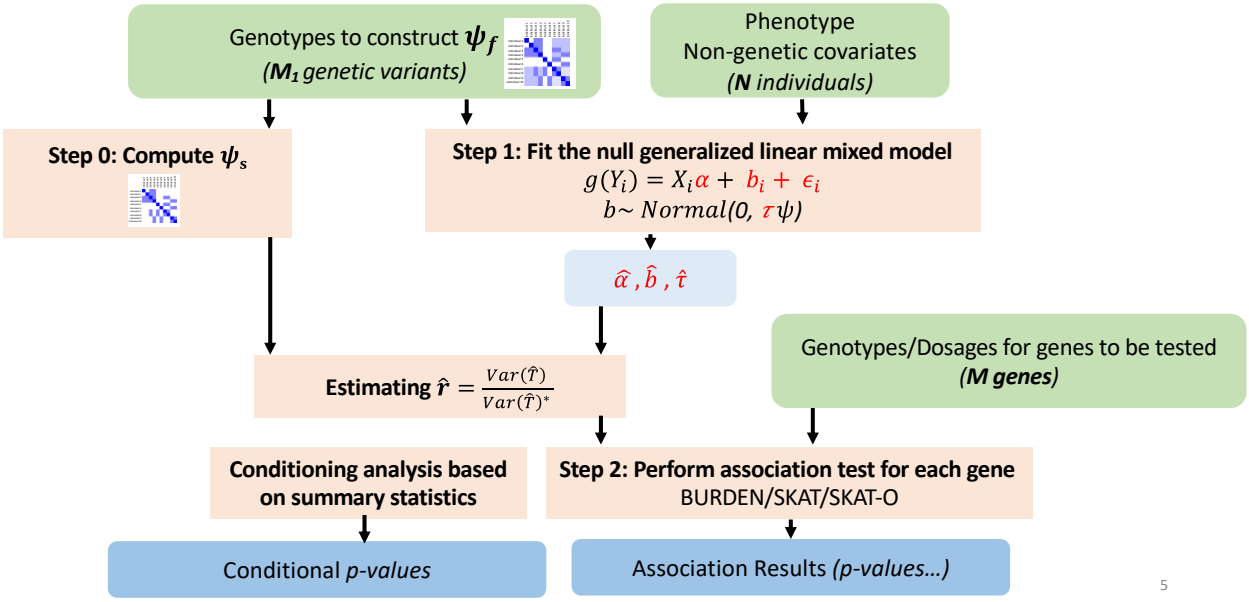
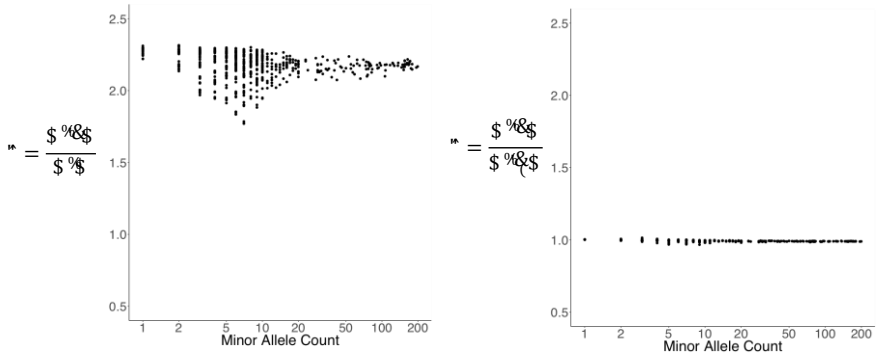
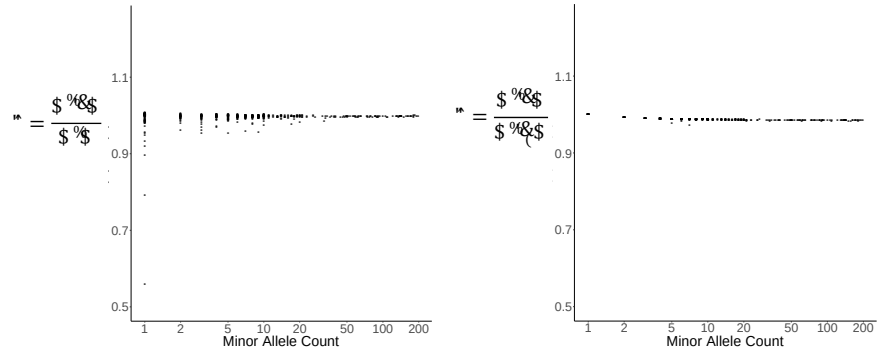


Figure S3.2. Plots of the variance ratio of the score statistics by MAC for rare variants with and without the full GRM for sample relatedness (left) and with the full GRM and a sparse GRM for closely related samples(right). A. 500 families and 5,000 independent individuals were simulated with h^2= 0.2 based on the pedigree structure shown in **Figure S3.11.** The sparse GRM was constructed using a coefficient of relatedness cutoff 0.2. B. 20,000 samples with White British ancestry were randomly selected from the UK Biobank and the null model was fitted for the automated read pulse rate. The sparse GRM was constructed using a coefficient of relatedness cutoff 0.125. C. 20,000 samples were randomly selected form the HUNT study and the null model was fitted for HDL. The sparse GRM was constructed using a coefficient of relatedness cutoff 0.125.

A. Simulation: 500 families and 5,000 independent individuals



B. UK Biobank: Pulse rate automated read (mean), randomly selected 20,000 individuals



C. HUNT: HDL, randomly selected 20,000 individuals

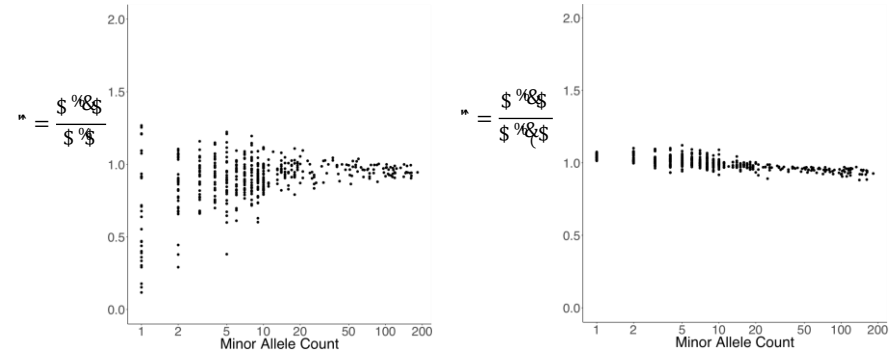
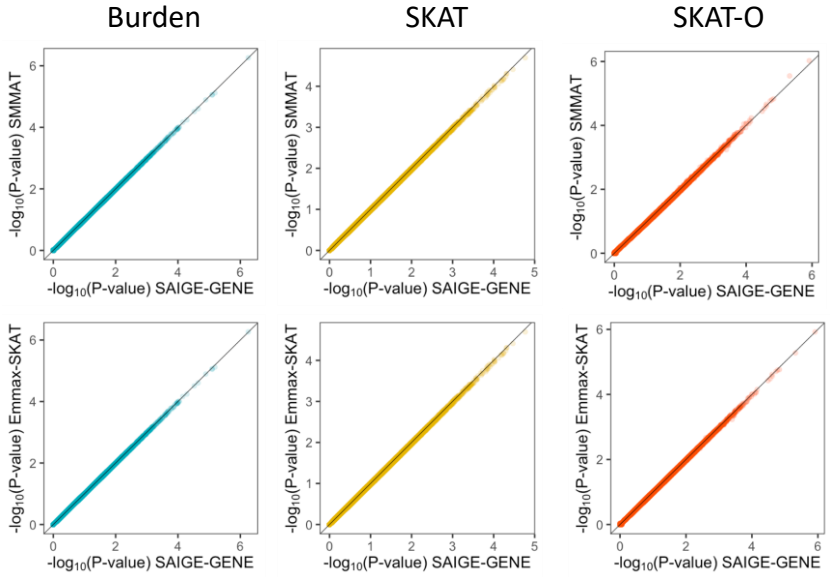


Figure S3.3. Scatter plots of association p-values from SAIGE-GENE versus SMMAT(Chen et al., 2018) and EmmaX-SKAT for the Burden, SKAT, and SKAT-O tests based on simulation data on the -log10 scale. 1,000,000 genes were tested with 1000 families, each having 10 members, as shown in the **Figure S3.11.** The Pearson's correlation coefficients $r^2 > 0.99$ for -log10(P-values) between SAIGE and SMMAT and between SAIGE and EmmaX-SKAT. A. $h^2 = 0.2$, B. $h^2 = 0.4$

A. $h^2 = 0.2$



B. $h^2 = 0.4$

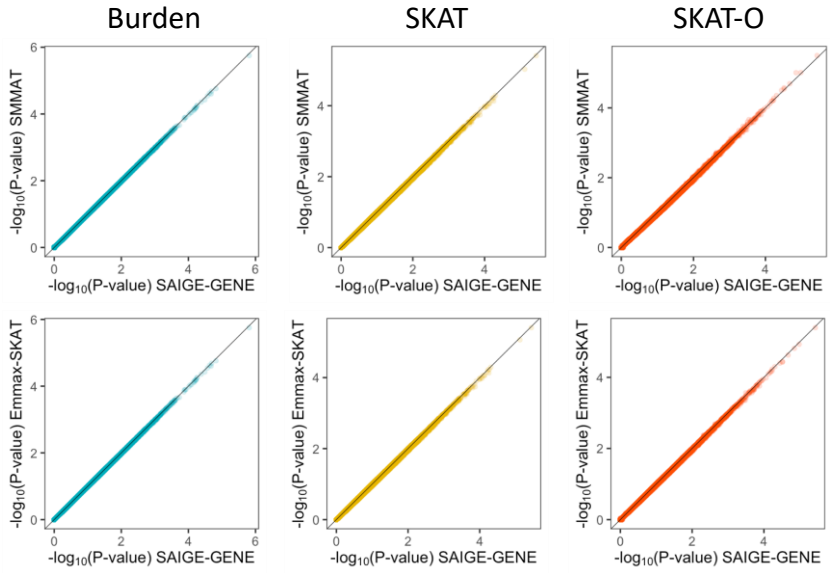
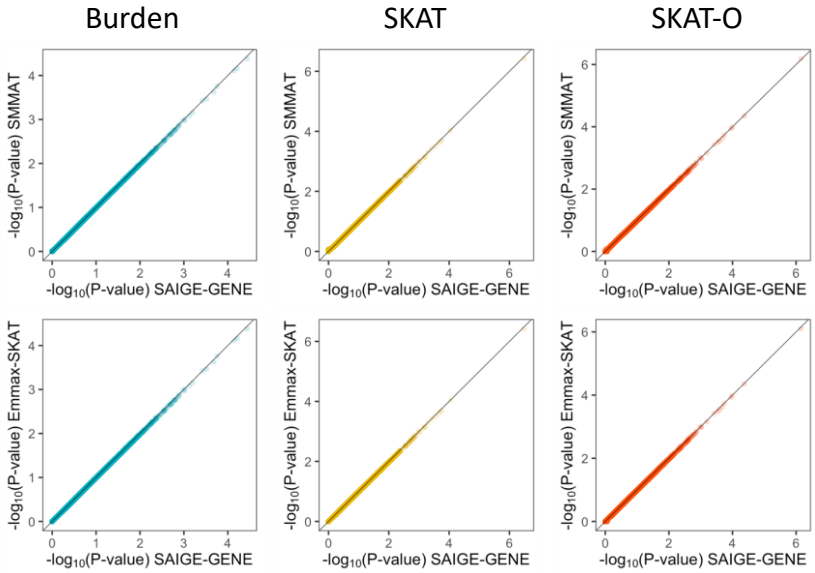


Figure S3.4. Scatter plots of association p-values from SAIGE-GENE versus SMMAT and EmmaX-SKAT for the Burden, SKAT, and SKAT-O tests based on real data analysis on the -log10 scale. 12,000 genes were tested for A. automated read pulse rate using 20,000 randomly selected white British samples in the HRC-imputed UK Biobank; B. HDL using 20,000 randomly selected samples in HUNT. Missense and stop-gain variants with MAF ≤ 1% were included. The Pearson's correlation coefficients $r^2 > 0.99$ for -log10(P-values) between SAIGE and SMMAT and between SAIGE and EmmaX-SKAT.

A. automated read pulse rate in the UK Biobank



B. HDL in the HUNT study

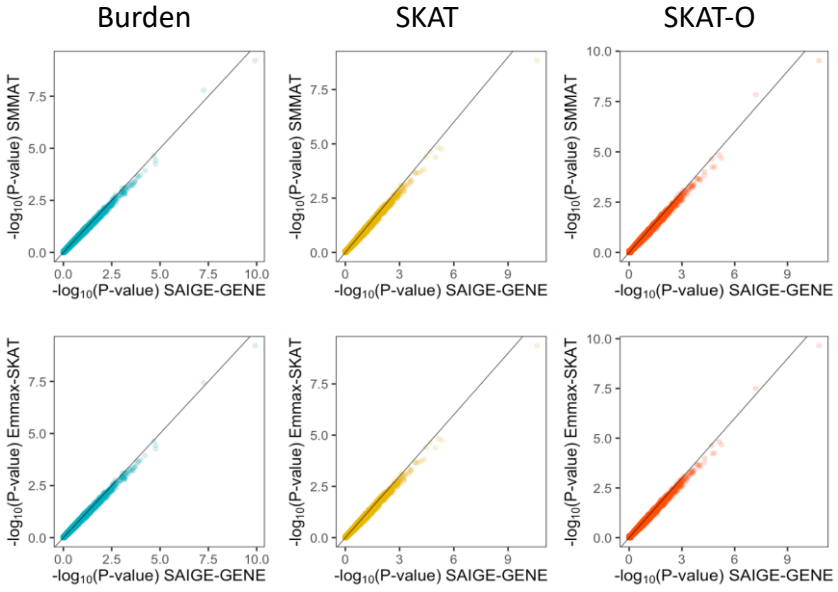


Figure S3.5. Scatter plots of association p-values on the -log10 scale from SAIGE-GENE with two sample relatedness cutoffs for the sparse GRM, 0.125 and 0.2. 15,338 genes were tested for automated read pulse rate in white British samples in the HRC-imputed UK Biobank. Missense and stop-gain variants with MAF ≤ 1% were included. A. Burden. B. SKAT, C. SKAT-O

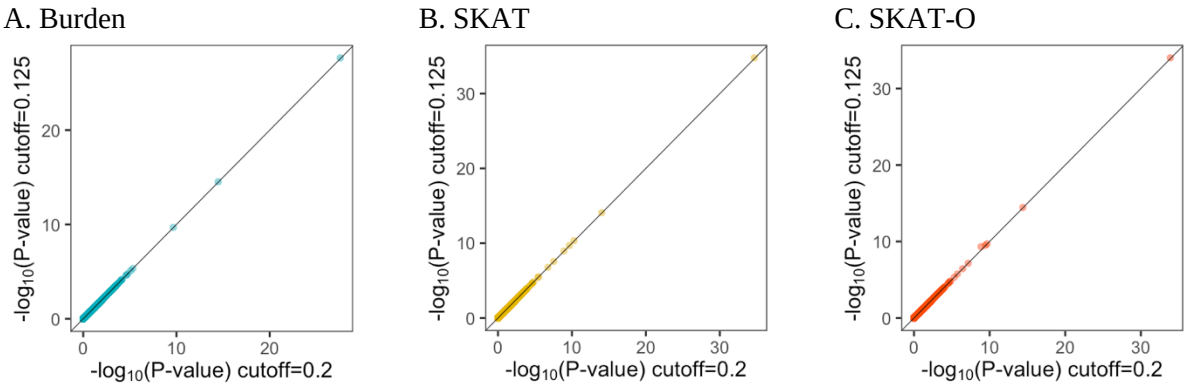


Figure S3.6. Quantile-quantile plots of association p-values for 10 million variant sets from the simulation study for phenotypes with various case-control ratios.

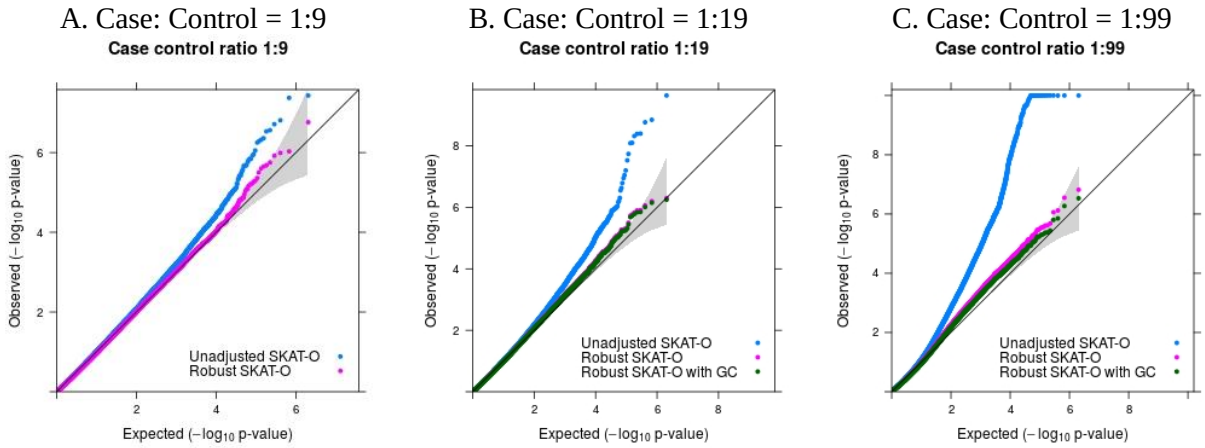


Figure S3.7. Empirical computation time for A. step 1 for fitting a null mixed model and B. step 2 for association tests, respectively by sample sizes (N) for gene-based tests for 15,342 genes, each containing 50 rare variants. Benchmarking was performed on randomly sub-sampled UK Biobank data with 408,144 White British participants for waist-to-hip ratio. The reported run times and memory are medians of five runs with samples randomly selected from the full sample set using different sampling seeds. The reported computation time and memory for EmmaX-SKAT and SMMAT is the projected computation time when N > 20,000. As the number of tested markers varies by sample sizes, the computation time is projected for 50 markers per gene for plotting. Numerical data are provided in **Table S3.1**.

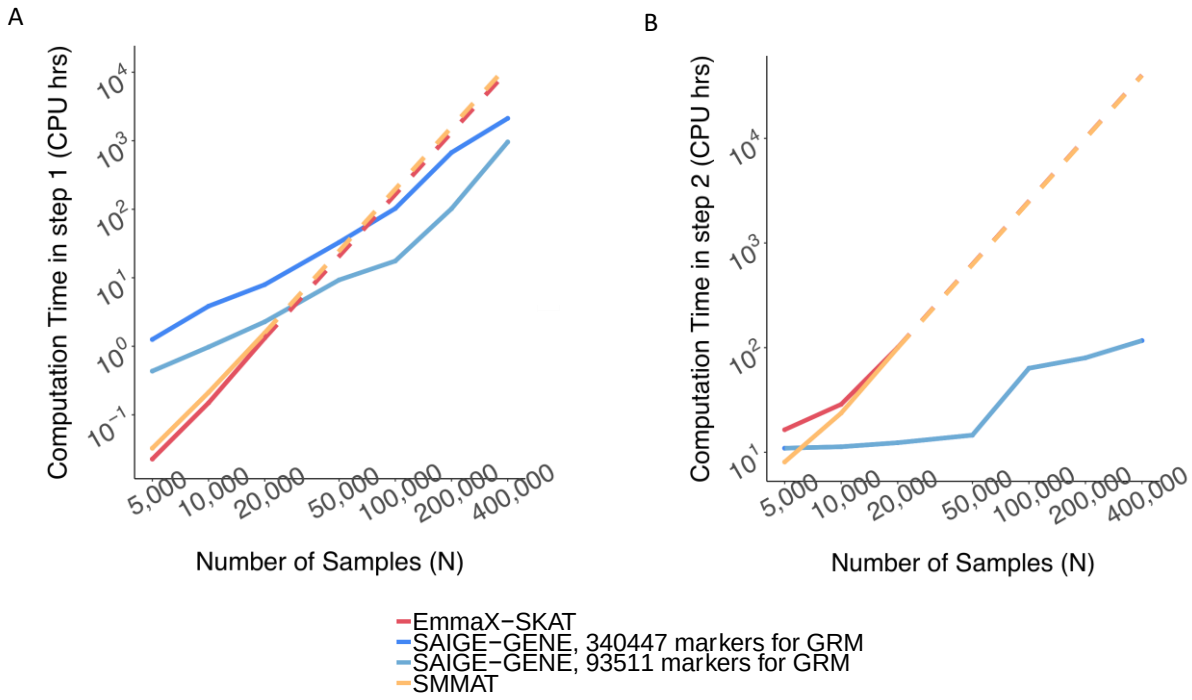


Figure S3.8. Log-log plot of the estimated run time as a function of number of markers per gene. Benchmarking was performed on randomly sub-sampled 400,000 UK Biobank data with 408,144 white British participants for waist-to-hip ratio on 15,342 genes. Run times are medians of five runs with samples randomly selected from the full sample set using different sampling seeds. The computation time for other different number of markers per gene is projected based on the benchmarked time.

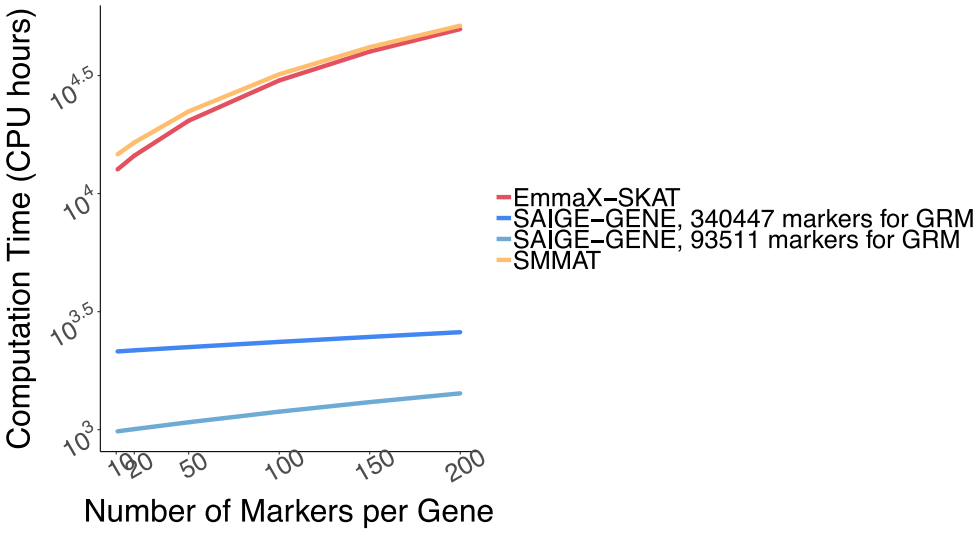


Figure S3.9. Log-log plots of the estimated A. run time and B. memory usage as a function of sample size (N) for genome-wide tests for 286,000 chunks, each containing 50 variants on average, given that there are 14.3 million markers in the HRC-imputed UK Biobank with MAF ≤ 1% and imputation info score ≥ 0.8. Numerical data are provided in **Table S3.1**. Benchmarking was performed on randomly sub-sampled UK Biobank data with 408,144 white British participants for waist-to-hip ratio. run times are medians of five runs with samples randomly selected from the full sample set using different sampling seeds.

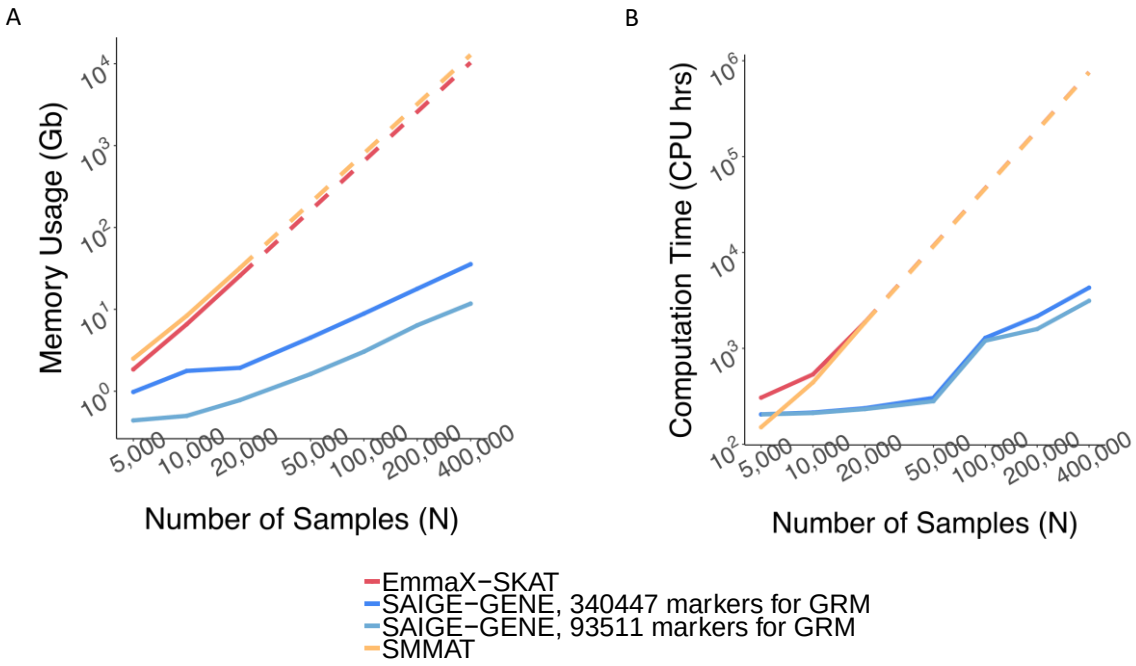


Figure S3.10. Log-log plots of the estimated run time for as a function of sample size (N) for SAIGE-GENE with and without using the robust adjustment. A. Exome-wide gene-based tests for 15,871 genes and B. Genome-wide tests for 286,000 chunks. Each gene or chunk contains 50 variants on average. Benchmarking was performed on randomly sub-sampled UK Biobank data with 402,163 white British participants tested for glaucoma (PheCode: 365, 4,462 cases and 397,701 controls). The case-control ratio remained the same in subsampled data sets. The reported run times and memory are medians of five runs with samples randomly selected from the full sample set using different sampling seeds. As the number of tested markers varies by sample sizes, the computation time is projected for 50 markers per gene for plotting. Numerical data are provided in **Table S3.2**.

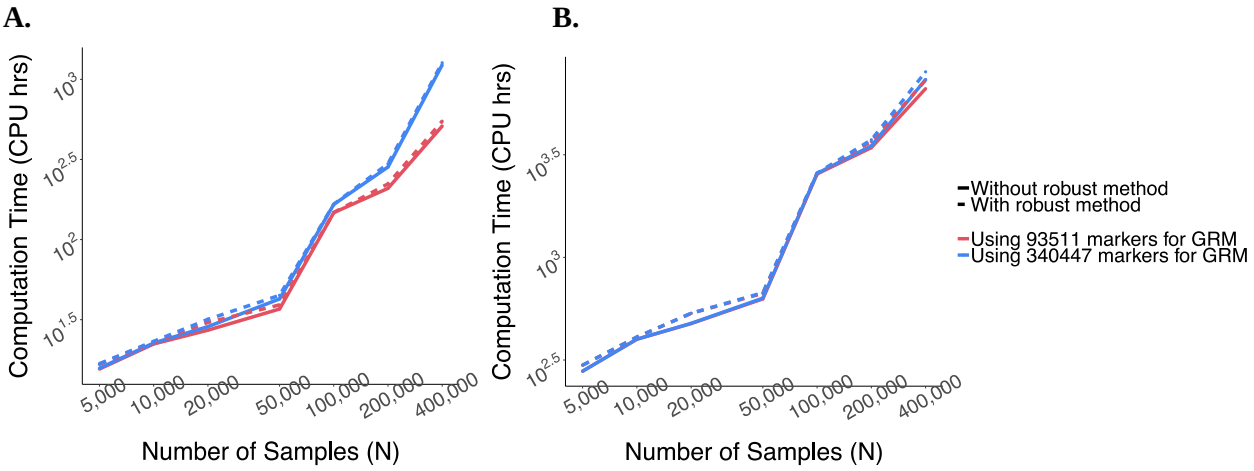


Figure S3.11. Pedigree of families, each with 10 members, in the simulation study.

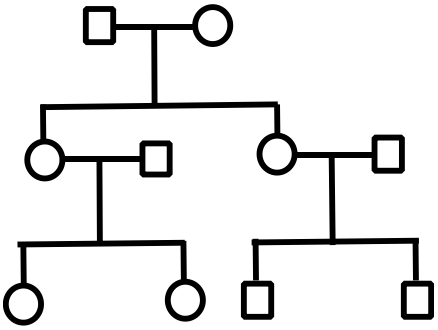
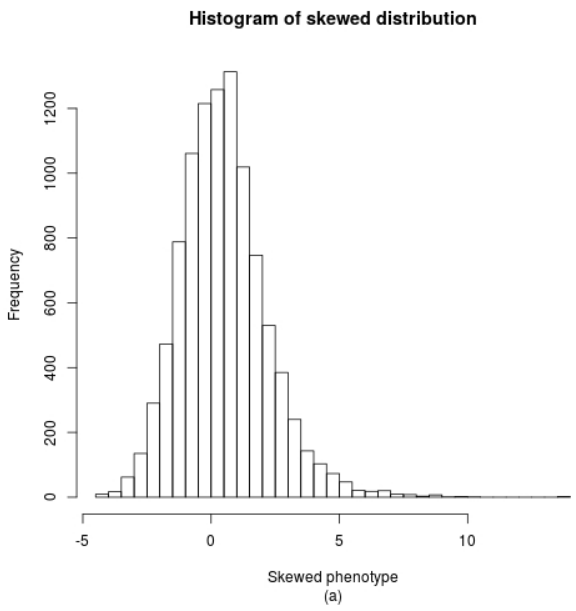


Figure S3.12. Histogram of simulated phenotypes. A. with skewed distribution. B. after inverse normal transformation as described in the subsection 2.2 of the section "Additional simulation and real-data analysis results"

A.



B.

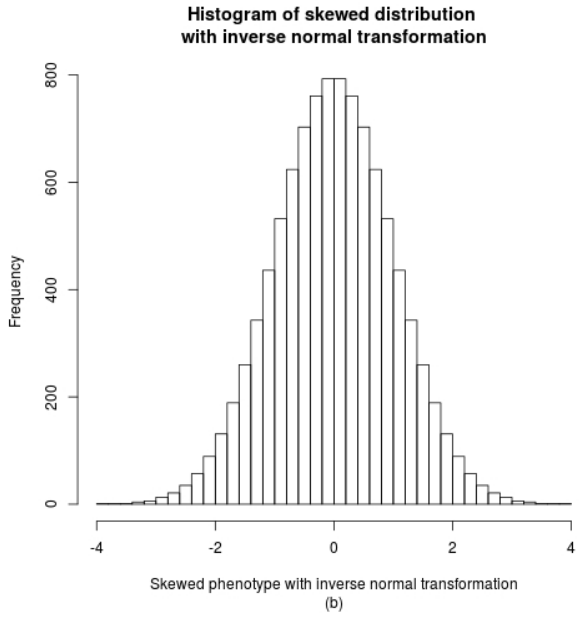
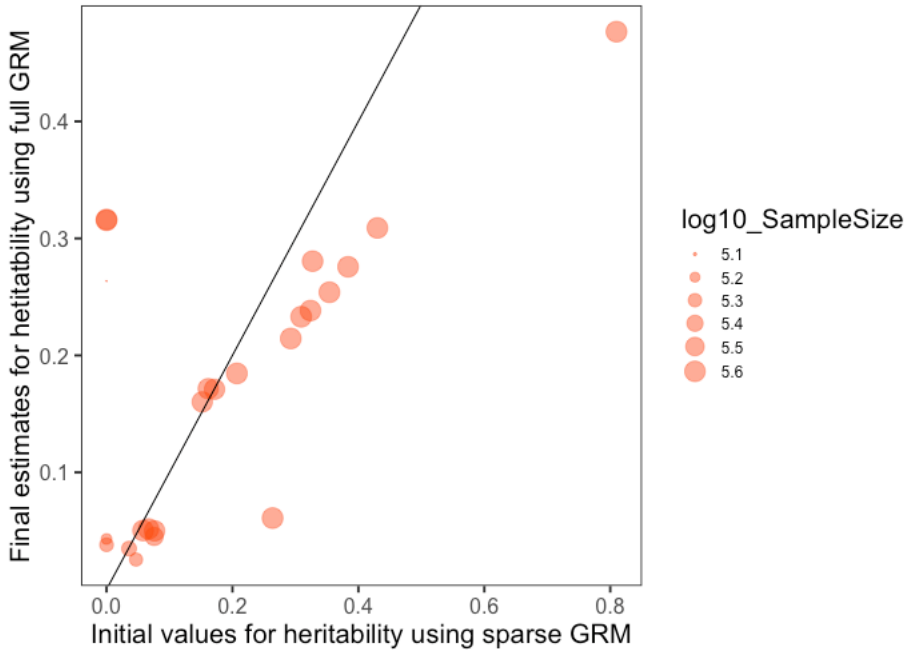


Figure S3.13. Comparing heritability estimates using the sparse GRM to heritability estimates using the full GRM for 24 quantitative traits in the UK Biobank with sample size (N) ≥ 100,000. The sparse GRM was constructed using a coefficient of relatedness cutoff 0.125, corresponding to up to 3rd degree relative pairs.



3.5.4 Supplementary tables

Table S3.1. The estimated run time (A) and memory use (B) across different sample sizes. Benchmarking was performed on randomly sub-sampled UK Biobank data with 408,144 white British participants and 15,342 genes were tested for waist hip ratio. For simplicity, the number of markers in the gene was 50 regardless of sample sizes. The run times and memory are medians of five runs with samples randomly selected from the full sample set using different sampling seeds. The computation cost for genome-wide region-based tests were projected from the exome-wide gene-based tests results given that there are 14.3 million markers in the HRC-imputed UK Biobank with MAF $\leq$ 1% and imputation info score $\geq$ 0.8. Total 286,000 chunks are tested with 50 markers in each chunk.

	sampleSize	step1(CPU hrs)	step2(CPU hrs)	Total Time(CPU hrs)	Memory (Gb)	Program
Exome-wide gene-based tests	5,000	0.43	10.94	11.37	0.44	SAIGE-GENE; 93,511 markers for GRM
	10,000	0.98	11.29	12.26	0.50	SAIGE-GENE; 93,511 markers for GRM
	20,000	2.28	12.34	14.61	0.78	SAIGE-GENE; 93,511 markers for GRM
	50,000	9.29	14.57	23.87	1.62	SAIGE-GENE; 93,511 markers for GRM
	100,000	17.58	63.51	81.09	3.04	SAIGE-GENE; 93,511 markers for GRM
	200,000	101.85	79.87	181.71	6.39	SAIGE-GENE; 93,511 markers for GRM
	400,000	958.98	116.65	1075.63	11.74	SAIGE-GENE; 93,511 markers for GRM
	5,000	1.26	10.94	12.20	0.98	SAIGE-GENE; 340,447 markers for GRM
	10,000	3.85	11.29	15.14	1.77	SAIGE-GENE; 340,447 markers for GRM
	20,000	7.95	12.34	20.28	1.93	SAIGE-GENE; 340,447 markers for GRM
	50,000	32.88	14.57	47.45	4.51	SAIGE-GENE; 340,447 markers for GRM
	100,000	102.71	63.51	166.22	8.87	SAIGE-GENE; 340,447 markers for GRM
	200,000	672.29	79.87	752.16	17.82	SAIGE-GENE; 340,447 markers for GRM
	400,000	2120.89	116.65	2237.54	35.59	SAIGE-GENE; 340,447 markers for GRM
	5,000	0.02	16.42	16.45	1.85	EmmaX-SKAT
	10,000	0.15	28.72	28.87	6.57	EmmaX-SKAT
	20,000	1.32	100.32	101.64	25.82	EmmaX-SKAT
	50,000	20.62	626.98	647.60	161.37	EmmaX-SKAT
	100,000	164.93	2507.93	2672.86	645.50	EmmaX-SKAT
	200,000	1319.44	10031.73	11351.17	2581.99	EmmaX-SKAT
	400,000	10555.55	40126.91	50682.46	10327.96	EmmaX-SKAT
	5,000	0.03	8.07	8.11	2.50	SMMAT
	10,000	0.21	23.68	23.90	8.34	SMMAT
	20,000	1.57	99.44	101.01	32.03	SMMAT
	50,000	24.57	621.47	646.04	200.22	SMMAT
	100,000	196.56	2485.89	2682.46	800.87	SMMAT
	200,000	1572.50	9943.57	11516.07	3203.49	SMMAT
	400,000	12580.00	39774.28	52354.28	12813.95	SMMAT
Genome-wide region-based tests	5,000	0.43	203.89	204.33	0.44	SAIGE-GENE; 93,511 markers for GRM
	10,000	0.98	210.42	211.40	0.50	SAIGE-GENE; 93,511 markers for GRM
	20,000	2.28	229.96	232.24	0.78	SAIGE-GENE; 93,511 markers for GRM
	50,000	9.29	271.67	280.97	1.62	SAIGE-GENE; 93,511 markers for GRM
	100,000	17.58	1183.98	1201.56	3.04	SAIGE-GENE; 93,511 markers for GRM
	200,000	101.85	1488.86	1590.70	6.39	SAIGE-GENE; 93,511 markers for GRM
	400,000	958.98	2174.55	3133.53	11.74	SAIGE-GENE; 93,511 markers for GRM
	5,000	1.26	203.89	205.15	0.98	SAIGE-GENE; 340,447 markers for GRM

	10,000	3.85	210.42	214.27	1.77	SAIGE-GENE; 340,447 markers for GRM
	20,000	7.95	229.96	237.90	1.93	SAIGE-GENE; 340,447 markers for GRM
	50,000	32.88	271.67	304.55	4.51	SAIGE-GENE; 340,447 markers for GRM
	100,000	102.71	1183.98	1286.69	8.87	SAIGE-GENE; 340,447 markers for GRM
	200,000	672.29	1488.86	2161.15	17.82	SAIGE-GENE; 340,447 markers for GRM
	400,000	2120.89	2174.55	4295.44	35.59	SAIGE-GENE; 340,447 markers for GRM
	5,000	0.02	306.15	306.17	1.85	EmmaX-SKAT
	10,000	0.15	535.40	535.55	6.57	EmmaX-SKAT
	20,000	1.32	1870.08	1871.40	25.82	EmmaX-SKAT
	50,000	20.62	11687.99	11708.61	161.37	EmmaX-SKAT
	100,000	164.93	46751.96	46916.89	645.50	EmmaX-SKAT
	200,000	1319.44	187007.85	188327.29	2581.99	EmmaX-SKAT
	400,000	10555.55	748031.39	758586.94	10327.96	EmmaX-SKAT
	5,000	0.03	150.51	150.55	2.50	SMMAT
	10,000	0.21	441.50	441.71	8.34	SMMAT
	20,000	1.57	1853.64	1855.22	32.03	SMMAT
	50,000	24.57	11585.28	11609.85	200.22	SMMAT
	100,000	196.56	46341.12	46537.68	800.87	SMMAT
	200,000	1572.50	185364.46	186936.96	3203.49	SMMAT
	400,000	12580.00	741457.85	754037.85	12813.95	SMMAT

Table S3.2. The estimated run time (A) and memory use (B) across different sample sizes for binary traits with and without the robust adjustment. Benchmarking was performed on randomly sub-sampled UK Biobank data with 402,163 white British participants and 15,871 genes were tested for glaucoma (PheCode: 365). Samples were randomly selected from 4,462 glaucoma cases and 397,701 controls respectively, so the case-control ratio remained the same in sub-sampled data sets. For simplicity, the number of markers in the gene was 50 regardless of sample sizes. The run times and memory are medians of five runs with samples randomly selected from the full sample set using different sampling seeds. The computation cost for genome-wide region-based tests were projected from the exome-wide gene-based tests results given that there are 14.3 million markers in the HRC-imputed UK Biobank with MAF $\leq$ 1% and imputation info score $\geq$ 0.8. Total 286,000 chunks are tested with 50 markers in each chunk.

	sampleSize	step1(CPU hrs)	step2(CPU hrs)	Total Time(CPU hrs)	Memory (Gb)	Program
Exome-wide gene-based tests	5,000	0.09	15.47	15.56	0.35	SAIGE-GENE without robust adjustment; 93,511 markers for GRM
	10,000	0.16	22.14	22.30	0.47	SAIGE-GENE without robust adjustment; 93,511 markers for GRM
	20,000	0.84	26.37	27.21	0.73	SAIGE-GENE without robust adjustment; 93,511 markers for GRM
	50,000	2.18	34.69	36.87	2.68	SAIGE-GENE without robust adjustment; 93,511 markers for GRM
	100,000	6.11	141.10	147.21	2.97	SAIGE-GENE without robust adjustment; 93,511 markers for GRM
	200,000	20.05	188.79	208.84	6.39	SAIGE-GENE without robust adjustment; 93,511 markers for GRM
	400,000	151.22	360.24	511.46	12.76	SAIGE-GENE without robust adjustment; 93,511 markers for GRM
	5,000	0.09	16.55	16.64	0.35	SAIGE-GENE with robust adjustment; 93,511 markers for GRM
	10,000	0.16	22.68	22.84	0.47	SAIGE-GENE with robust adjustment; 93,511 markers for GRM
	20,000	0.84	29.55	30.39	0.73	SAIGE-GENE with robust adjustment; 93,511 markers for GRM
	50,000	2.18	36.94	39.12	2.68	SAIGE-GENE with robust adjustment; 93,511 markers for GRM
	100,000	6.11	141.58	147.69	2.97	SAIGE-GENE with robust adjustment; 93,511 markers for GRM
	200,000	20.05	202.43	222.48	6.39	SAIGE-GENE with robust adjustment; 93,511 markers for GRM
	400,000	151.22	396.94	548.16	12.76	SAIGE-GENE with robust adjustment; 93,511 markers for GRM
	5,000	0.26	15.47	15.73	0.98	SAIGE-GENE without robust adjustment; 340,447 markers for GRM
	10,000	0.49	22.14	22.63	1.06	SAIGE-GENE without robust adjustment; 340,447 markers for GRM
	20,000	2.24	26.37	28.61	3.39	SAIGE-GENE without robust adjustment; 340,447 markers for GRM
	50,000	7.93	34.69	42.62	4.44	SAIGE-GENE without robust adjustment; 340,447 markers for GRM
	100,000	24.68	141.10	165.79	8.94	SAIGE-GENE without robust adjustment; 340,447 markers for GRM
	200,000	94.84	188.79	283.63	17.51	SAIGE-GENE without robust adjustment; 340,447 markers for GRM
	400,000	872.30	360.24	1232.54	35.38	SAIGE-GENE without robust adjustment; 340,447 markers for GRM
	5,000	0.26	16.55	16.81	0.98	SAIGE-GENE with robust adjustment; 340,447 markers for GRM
	10,000	0.49	22.68	23.18	1.06	SAIGE-GENE with robust adjustment; 340,447 markers for GRM
	20,000	2.24	29.55	31.79	3.39	SAIGE-GENE with robust adjustment; 340,447 markers for GRM
	50,000	7.93	36.94	44.87	4.44	SAIGE-GENE with robust adjustment; 340,447 markers for GRM
	100,000	24.68	141.58	166.27	8.94	SAIGE-GENE with robust adjustment; 340,447 markers for GRM
	200,000	94.84	202.43	297.26	17.51	SAIGE-GENE with robust adjustment; 340,447 markers for GRM
	400,000	872.30	396.94	1269.25	35.38	SAIGE-GENE with robust adjustment; 340,447 markers for GRM
Genome-wide region-based tests	5,000	0.09	278.79	278.88	0.35	SAIGE-GENE without robust adjustment; 93,511 markers for GRM
	10,000	0.16	398.98	399.14	0.47	SAIGE-GENE without robust adjustment; 93,511 markers for GRM
	20,000	0.84	475.16	476.00	0.73	SAIGE-GENE without robust adjustment; 93,511 markers for GRM
	50,000	2.18	625.10	627.28	2.68	SAIGE-GENE without robust adjustment; 93,511 markers for GRM
	100,000	6.11	2542.81	2548.91	2.97	SAIGE-GENE without robust adjustment; 93,511 markers for GRM
	200,000	20.05	3402.16	3422.22	6.39	SAIGE-GENE without robust adjustment; 93,511 markers for GRM
	400,000	151.22	6491.80	6643.02	12.76	SAIGE-GENE without robust adjustment; 93,511 markers for GRM
	5,000	0.09	298.30	298.39	0.35	SAIGE-GENE with robust adjustment; 93,511 markers for GRM
	10,000	0.16	408.80	408.96	0.47	SAIGE-GENE with robust adjustment; 93,511 markers for GRM

	20,000	0.84	532.50	533.34	0.73	SAIGE-GENE with robust adjustment; 93,511 markers for GRM
	50,000	2.18	665.67	667.85	2.68	SAIGE-GENE with robust adjustment; 93,511 markers for GRM
	100,000	6.11	2551.42	2557.53	2.97	SAIGE-GENE with robust adjustment; 93,511 markers for GRM
	200,000	20.05	3647.87	3667.92	6.39	SAIGE-GENE with robust adjustment; 93,511 markers for GRM
	400,000	151.22	7153.23	7304.45	12.76	SAIGE-GENE with robust adjustment; 93,511 markers for GRM
	5,000	0.26	278.79	279.05	0.98	SAIGE-GENE without robust adjustment; 340,447 markers for GRM
	10,000	0.49	398.98	399.47	1.06	SAIGE-GENE without robust adjustment; 340,447 markers for GRM
	20,000	2.24	475.16	477.40	3.39	SAIGE-GENE without robust adjustment; 340,447 markers for GRM
	50,000	7.93	625.10	633.03	4.44	SAIGE-GENE without robust adjustment; 340,447 markers for GRM
	100,000	24.68	2542.81	2567.49	8.94	SAIGE-GENE without robust adjustment; 340,447 markers for GRM
	200,000	94.84	3402.16	3497.00	17.51	SAIGE-GENE without robust adjustment; 340,447 markers for GRM
	400,000	872.30	6491.80	7364.11	35.38	SAIGE-GENE without robust adjustment; 340,447 markers for GRM
	5,000	0.26	298.30	298.56	0.98	SAIGE-GENE with robust adjustment; 340,447 markers for GRM
	10,000	0.49	408.80	409.29	1.06	SAIGE-GENE with robust adjustment; 340,447 markers for GRM
	20,000	2.24	532.50	534.75	3.39	SAIGE-GENE with robust adjustment; 340,447 markers for GRM
	50,000	7.93	665.67	673.60	4.44	SAIGE-GENE with robust adjustment; 340,447 markers for GRM
	100,000	24.68	2551.42	2576.10	8.94	SAIGE-GENE with robust adjustment; 340,447 markers for GRM
	200,000	94.84	3647.87	3742.70	17.51	SAIGE-GENE with robust adjustment; 340,447 markers for GRM
	400,000	872.30	7153.23	8025.54	35.38	SAIGE-GENE with robust adjustment; 340,447 markers for GRM

Table S3.3. Heritability estimated based on the full GRM by step 1 in SAIGE-GENE A. for 53 quantitative traits and B. for 10 binary traits in the UK Biobank. For quantitative traits, $\hat{h}^2 = \hat{\tau}/(\hat{\phi} + \hat{\tau})$, where $\hat{\tau}$ is the additive genetic variance parameter estimate and $\hat{\phi}$ is the variance parameter estimate for the error term (see **ONLINE METHODS**) from step 1 in SAIGE-GENE. For binary traits, heritability estimates in a liability scale, $\hat{h}^2_{latent}$, is reported here. The $\hat{h}^2_{latent}$ was obtained using the fact that the logistic regression can be described as a liability threshold model with standard logistic distribution, which has variance=$\pi^2/3$ = 3.23. $\hat{h}^2_{latent} = \frac{\hat{\tau}}{\pi^2/3+\hat{\tau}}$, where $\hat{\tau}$ is the additive genetic variance parameter estimate from step 1 in SAIGE-GENE. Since GRM was constructed using variants with MAF ≥ 0.01, the estimated heritability is the narrow sense of the heritability due to common variants.

A.

Phenotype	**Sample Size**	$\hat{h}^2$
Townsend_deprivation_index	408422	0.061
Waist_circumference	408227	0.214
Hip_circumference	408182	0.233
Waist_hip_ratio	408144	0.185
Height_standing	408034	0.477
Body_mass_index	407605	0.254
Days_per_week_walked_10min	402016	0.050
Weight	401786	0.276
Whole_body_water_mass	401782	0.315
Basal_metabolic_rate	401771	0.309
Whole_body_fat_free_mass	401747	0.316
Body_fat_percentage	401556	0.238
Days_week_vigorous_phys_activity_10min	389393	0.050
Days_per_week_moderate_phys_activity_10min	389204	0.052
Blood_pressure_diastolic_automated_mean	385365	0.160
Pulse_rate_automated_mean	385365	0.171
Blood_pressure_systolic_automated_mean	385362	0.172
Duration_moderate_activity	303663	0.045
Duration_of_vigorous_activity	221867	0.035
Duration_of_light_DIY	203191	0.038
Duration_of_other_exercises	194610	0.025
Duration_of_heavy_DIY	165319	0.043
Smoking_packyear	124011	0.264
Age_high_blood_pressure	99296	0.106
Duration_of_strenuous_sports	41133	0.032
Blood_pressure_diastolic_manual_mean	35283	0.149
Blood_pressure_systolic_manual_mean	35283	0.156
Age_diabetes	18552	0.187
Age_at_death	11928	0.000
Heart_rate	8421	0.208
P_wave_duration	8421	0.296
QRS_duration	8421	0.337
Average_heart_rate_MRI	3875	0.231
Body_surface_area_MRI	3875	0.521
Cardiac_index	3875	0.083
Cardiac_output	3875	0.154
LV_ejection_fraction	3875	0.504
LV_end_diastolic_volume	3875	0.650
LV_end_systolic_volume	3875	0.522
LV_stroke_volume	3875	0.681

Maximum_carotid_IMT_150	2221	0.135
Mean_carotid_IMT_150	2221	0.379
Minimum_carotid_IMT_150	2221	0.590
Maximum_carotid_IMT_210	2209	0.830
Mean_carotid_IMT_210	2209	0.790
Minimum_carotid_IMT_210	2209	0.678
Maximum_carotid_IMT_120	2204	0.000
Mean_carotid_IMT_120	2204	0.000
Minimum_carotid_IMT_120	2204	0.000
Maximum_carotid_IMT_240	2183	0.178
Mean_carotid_IMT_240	2183	0.300
Minimum_carotid_IMT_240	2183	0.000

B.

Phenotype	**Phecode**	**Number of cases**	**Number of controls**	$\widehat{h}^2_{latent}$
Hematuria	593	16409	379936	0.074
Cholelithiasis and cholecystitis	574	16225	391307	0.121
Prostate cancer	185	6743	169185	0.082
Diseases of hair and hair follicles	704	5344	402357	0.156
Colorectal cancer	153	4562	382756	0.135
Glaucoma	365	4462	397761	0.133
Pulmonary heart disease	415	4257	402375	0.111
Melanomas of skin, dx or hx	172.1	2691	395071	0.161
Ankylosing spondylitis	715.2	620	365085	0.000
Thyroid cancer	193	358	407399	0.000

Table S3.4. Exome-wide significant genes with p-values ≤ $2.5x10^{-6}$ identified by SAIGE-GENE in the UK Biobank a. for 53 quantitative traits, b for 10 binary traits with various case-control ratios.

A.

Phenotype	Gene	P-value	Number of variants	Sample Size
Waist circumference	*GPR151*	2.15E-10	6	408227
Waist circumference	*C16orf70*	1.94E-06	2	408227
Hip circumference	*SYPL2*	2.91E-08	4	408182
Hip circumference	*TRAPPC4*	5.83E-07	2	408182
Hip circumference	*ANO1*	5.98E-07	4	408182
Hip circumference	*GPR151*	6.08E-07	6	408182
Hip circumference	*C16orf70*	1.14E-06	2	408182
Waist hip ratio	*TAS2R46*	1.36E-08	2	408144
Waist hip ratio	*GPR151*	3.00E-08	6	408144
Waist hip ratio	*SLC5A3*	1.33E-07	4	408144
Height standing	*SCMH1*	3.41E-39	8	408034
Height standing	*ACAN*	1.39E-36	33	408034
Height standing	*FBN2*	3.15E-32	16	408034
Height standing	*ZFAT*	3.61E-32	13	408034
Height standing	*HTRA1*	1.66E-29	4	408034
Height standing	*NPR3*	3.60E-25	3	408034
Height standing	*STC2*	1.10E-24	4	408034
Height standing	*SPSB3*	1.10E-22	4	408034
Height standing	*NUBP2*	2.01E-22	9	408034
Height standing	*ATAD2*	8.07E-21	6	408034
Height standing	*ADAMTS6*	1.83E-17	6	408034
Height standing	*GRAMD2A*	3.41E-16	4	408034
Height standing	*GRM4*	7.23E-16	2	408034
Height standing	*MTMR11*	1.94E-15	3	408034
Height standing	*CRISPLD2*	2.04E-14	11	408034
Height standing	*CERCAM*	7.05E-14	10	408034
Height standing	*PDE3B*	1.88E-13	8	408034
Height standing	*FNDC3B*	1.39E-12	8	408034
Height standing	*PKD1*	6.40E-12	43	408034
Height standing	*FER*	6.68E-12	5	408034
Height standing	*C16orf70*	7.45E-12	2	408034
Height standing	*HAPLN3*	1.25E-11	6	408034
Height standing	*ST3GAL4*	5.37E-11	5	408034
Height standing	*SHANK1*	8.09E-11	5	408034
Height standing	*TRAPPC13*	1.04E-10	2	408034
Height standing	*S1PR5*	1.44E-10	5	408034
Height standing	*PTCH1*	2.65E-10	15	408034
Height standing	*COL8A1*	3.25E-10	2	408034
Height standing	*EXD1*	3.84E-10	3	408034
Height standing	*ATAD5*	4.74E-10	8	408034
Height standing	*ESR1*	7.28E-10	9	408034
Height standing	*CLEC3A*	9.08E-10	3	408034
Height standing	*PTH1R*	1.40E-09	3	408034
Height standing	*FGFR3*	3.50E-09	4	408034
Height standing	*NOX4*	4.09E-09	4	408034
Height standing	*CYR61*	4.40E-09	2	408034
Height standing	*TBX3*	4.59E-09	2	408034
Height standing	*SAMD4A*	4.77E-09	7	408034
Height standing	*ZCCHC6*	1.42E-08	7	408034
Height standing	*CTU2*	2.09E-08	12	408034
Height standing	*LPP*	2.11E-08	8	408034
Height standing	*LRRC8A*	2.87E-08	2	408034
Height standing	*DGKH*	3.08E-08	9	408034
Height standing	*ABCB6*	4.74E-08	11	408034

Height standing	*PIEZO1*	6.63E-08	55	408034
Height standing	*ELN*	6.67E-08	8	408034
Height standing	*ATG10*	1.35E-07	4	408034
Height standing	*CADM1*	1.35E-07	2	408034
Height standing	*CPPED1*	1.40E-07	7	408034
Height standing	*PFDN2*	1.41E-07	2	408034
Height standing	*PROP1*	1.63E-07	4	408034
Height standing	*PRSS56*	1.68E-07	3	408034
Height standing	*GYG1*	1.78E-07	5	408034
Height standing	*CHGA*	2.20E-07	9	408034
Height standing	*SERPINE2*	2.80E-07	3	408034
Height standing	*SPATA5*	2.83E-07	5	408034
Height standing	*AMOTL1*	2.85E-07	6	408034
Height standing	*TMEM150B*	2.88E-07	3	408034
Height standing	*COL11A1*	2.91E-07	8	408034
Height standing	*PTK7*	2.95E-07	6	408034
Height standing	*PHC3*	3.24E-07	7	408034
Height standing	*TARS2*	4.22E-07	3	408034
Height standing	*VASN*	5.25E-07	7	408034
Height standing	*TXLNA*	5.66E-07	3	408034
Height standing	*FAM76A*	5.70E-07	3	408034
Height standing	*C11orf57*	5.71E-07	2	408034
Height standing	*FBLN2*	6.61E-07	13	408034
Height standing	*MC3R*	6.92E-07	2	408034
Height standing	*GTF2E2*	7.05E-07	3	408034
Height standing	*FIBIN*	1.09E-06	2	408034
Height standing	*PREB*	1.12E-06	5	408034
Height standing	*SIX6*	1.16E-06	3	408034
Height standing	*DOT1L*	1.20E-06	5	408034
Height standing	*SETD2*	1.30E-06	19	408034
Height standing	*USP37*	1.40E-06	8	408034
Height standing	*BCKDHA*	1.91E-06	2	408034
Height standing	*HR*	1.92E-06	21	408034
Height standing	*PAM*	1.98E-06	5	408034
Height standing	*PLG*	2.04E-06	11	408034
Height standing	*LLGL1*	2.30E-06	9	408034
Height standing	*ZNF205*	2.42E-06	4	408034
Body mass index	*GPR151*	1.74E-09	6	407605
Body mass index	*SYPL2*	6.43E-09	4	407605
Body mass index	*TRAPPC4*	2.87E-07	2	407605
Body mass index	*IQSEC1*	7.02E-07	12	407605
Body mass index	*HCRTR2*	1.50E-06	4	407605
Body mass index	*FRMD5*	1.92E-06	4	407605
Weight	*ZFAT*	1.80E-11	13	401786
Weight	*C16orf70*	7.28E-11	2	401786
Weight	*STC2*	1.69E-09	4	401786
Weight	*GPR151*	1.73E-09	6	401786
Weight	*ANO1*	3.10E-08	4	401786
Weight	*SYPL2*	6.22E-08	4	401786
Weight	*FRMD5*	6.24E-08	4	401786
Weight	*NUBP2*	1.70E-07	9	401786
Weight	*SCMH1*	1.87E-07	8	401786
Weight	*TRAPPC4*	1.16E-06	2	401786
Weight	*HTRA1*	2.00E-06	4	401786
Weight	*SPSB3*	2.32E-06	4	401786
Whole body water mass	*STC2*	8.55E-24	4	401782
Whole body water mass	*NUBP2*	1.52E-19	9	401782
Whole body water mass	*ZFAT*	1.83E-18	13	401782
Whole body water mass	*SCMH1*	8.44E-18	8	401782
Whole body water mass	*SPSB3*	2.40E-17	4	401782
Whole body water mass	*HTRA1*	2.11E-12	4	401782
Whole body water mass	*ANO1*	4.31E-12	4	401782

Whole body water mass	*PHC3*	1.32E-11	7	401782
Whole body water mass	*C16orf70*	3.33E-11	2	401782
Whole body water mass	*ATAD2*	2.82E-09	6	401782
Whole body water mass	*GRM4*	6.47E-09	2	401782
Whole body water mass	*ESR1*	1.15E-07	9	401782
Whole body water mass	*ZMYM6*	1.75E-07	11	401782
Whole body water mass	*TMEM150B*	2.96E-07	3	401782
Whole body water mass	*FRMD5*	2.98E-07	4	401782
Whole body water mass	*LCOR*	3.88E-07	11	401782
Whole body water mass	*PLEKHJ1*	9.18E-07	5	401782
Whole body water mass	*GLI3*	1.09E-06	8	401782
Whole body water mass	*ACAN*	2.12E-06	33	401782
Basal metabolic rate	*STC2*	1.50E-20	4	401771
Basal metabolic rate	*ZFAT*	2.39E-17	13	401771
Basal metabolic rate	*NUBP2*	4.62E-17	9	401771
Basal metabolic rate	*SCMH1*	3.15E-15	8	401771
Basal metabolic rate	*SPSB3*	5.37E-15	4	401771
Basal metabolic rate	*C16orf70*	9.08E-12	2	401771
Basal metabolic rate	*HTRA1*	4.87E-11	4	401771
Basal metabolic rate	*ANO1*	5.53E-11	4	401771
Basal metabolic rate	*PHC3*	2.11E-10	7	401771
Basal metabolic rate	*GRM4*	9.53E-10	2	401771
Basal metabolic rate	*FRMD5*	1.35E-07	4	401771
Basal metabolic rate	*ATAD2*	2.27E-07	6	401771
Basal metabolic rate	*LCOR*	7.41E-07	11	401771
Basal metabolic rate	*ZMYM6*	8.04E-07	11	401771
Basal metabolic rate	*TMEM150B*	1.10E-06	3	401771
Basal metabolic rate	*ESR1*	1.14E-06	9	401771
Basal metabolic rate	*GPR151*	1.97E-06	6	401771
Whole body fat free mass	*STC2*	5.52E-24	4	401747
Whole body fat free mass	*NUBP2*	1.70E-19	9	401747
Whole body fat free mass	*ZFAT*	5.30E-19	13	401747
Whole body fat free mass	*SCMH1*	2.90E-18	8	401747
Whole body fat free mass	*SPSB3*	2.51E-17	4	401747
Whole body fat free mass	*HTRA1*	1.32E-12	4	401747
Whole body fat free mass	*ANO1*	6.36E-12	4	401747
Whole body fat free mass	*PHC3*	7.42E-12	7	401747
Whole body fat free mass	*C16orf70*	1.18E-10	2	401747
Whole body fat free mass	*GRM4*	6.51E-09	2	401747
Whole body fat free mass	*ATAD2*	6.80E-09	6	401747
Whole body fat free mass	*ESR1*	7.69E-08	9	401747
Whole body fat free mass	*ZMYM6*	2.27E-07	11	401747
Whole body fat free mass	*TMEM150B*	5.78E-07	3	401747
Whole body fat free mass	*FRMD5*	5.88E-07	4	401747
Whole body fat free mass	*LCOR*	1.06E-06	11	401747
Whole body fat free mass	*ACAN*	1.31E-06	33	401747
Whole body fat free mass	*PLEKHJ1*	1.72E-06	5	401747
Whole body fat free mass	*GLI3*	1.94E-06	8	401747
Whole body fat free mass	*PAM*	2.01E-06	5	401747
Impedance of whole body	*CYR61*	3.81E-18	2	401746
Impedance of whole body	*POR*	2.90E-12	7	401746
Impedance of whole body	*ADAMTS3*	7.23E-12	8	401746
Impedance of whole body	*ANO1*	4.91E-11	4	401746
Impedance of whole body	*STC2*	1.61E-09	4	401746
Impedance of whole body	*ECM2*	1.95E-08	10	401746
Impedance of whole body	*ZNF469*	3.93E-08	53	401746
Impedance of whole body	*NUBP2*	3.02E-07	9	401746
Impedance of whole body	*FBN2*	3.59E-07	16	401746
Impedance of whole body	*FAM198A*	1.08E-06	7	401746
Impedance of whole body	*ZMYM6*	1.18E-06	11	401746
Impedance of whole body	*SNED1*	1.99E-06	9	401746
Body fat percentage	*CYR61*	4.99E-12	2	401556

Body fat percentage	*GPR151*	2.85E-10	6	401556
Body fat percentage	*SYPL2*	2.81E-07	4	401556
Body fat percentage	*C10orf35*	3.64E-07	2	401556
Days per week moderate phys activity 10min	*RAD51AP1*	3.87E-07	7	389204
Blood pressure diastolic automated mean	*DBH*	3.08E-14	12	385365
Blood pressure diastolic automated mean	*SLC9A3R2*	2.90E-11	5	385365
Blood pressure diastolic automated mean	*ARID1B*	1.08E-06	7	385365
Pulse rate automated mean	*TBX5*	9.69E-35	4	385365
Pulse rate automated mean	*MYH6*	3.61E-15	14	385365
Pulse rate automated mean	*TTN*	3.18E-10	368	385365
Pulse rate automated mean	*KIF1C*	4.78E-10	12	385365
Pulse rate automated mean	*ARHGEF40*	7.02E-08	7	385365
Pulse rate automated mean	*FNIP1*	3.58E-07	8	385365
Pulse rate automated mean	*DBH*	1.74E-06	12	385365
Blood pressure systolic automated mean	*SLC9A3R2*	2.36E-12	5	385362
Blood pressure systolic automated mean	*ZFAT*	7.06E-12	13	385362
Blood pressure systolic automated mean	*DBH*	2.85E-10	12	385362
Blood pressure systolic automated mean	*RRAS*	1.33E-07	2	385362
Blood pressure systolic automated mean	*NOX4*	1.91E-07	4	385362
Blood pressure systolic automated mean	*TBX5*	2.87E-07	4	385362
Blood pressure systolic automated mean	*COL21A1*	8.05E-07	14	385362

B.

Phenotype	**Phecode**	**Gene**	**P-value**	**Number of variants**	**Number of cases**	**Number of controls**
Cholelithiasis and cholecystitis	574	ABCG5	2.31E-13	13	16225	391307
Cholelithiasis and cholecystitis	574	ABCG8	4.47E-10	12	16225	391307
Cholelithiasis and cholecystitis	574	ABCB4	6.86E-07	9	16225	391307
Sebaceous cyst	706.2	GORASP1	1.26E-18	8	8876	399255
Diseases of hair and hair follicles	704	GORASP1	4.36E-16	8	5344	402357
Glaucoma	365	MYOC	1.24E-06	6	4462	397761
Pulmonary heart disease	415	CREB3L1	9.59E-10	6	4257	402375
Ankylosing spondylitis	715.2	SLC44A4	1.23E-15	8	620	365085
Ankylosing spondylitis	715.2	PSMB9	5.83E-10	3	620	365085
Ankylosing spondylitis	715.2	IER3	1.86E-09	4	620	365085

Table S3.5. Exome-wide significant genes with p-values ≤ $2.5x10^{-6}$ identified by SAIGE-GENE but not identified by SAIGE in the UK Biobank for 53 quantitative traits

.

					Most significant variant in the locus (+/- 500kb of the start and end positions of the gene)							
Phenotype	**Gene**	**Number of Variants**	**Sample Size**	**P-value (SAKT-O)**	**chr:pos:ref:alt**	**Allele frequency**	**BETA**	**SE**	**P-value**	**In gene-based tests (0=no, 1=yes)**	**function**	**Gene**
Waist_circumference	*C16orf70*	2	408227	1.94E-06	16:66995927:G:A	0.00717746	0.07	0.01	5.57E-07	0	intronic	*CES3*
Hip_circumference	*GPR151*	6	408182	6.08E-07	5:146212195:C:G	3.14E-05	1.64	0.34	1.12E-06	0	intronic	*PPP2R2B*
Hip_circumference	*C16orf70*	2	408182	1.14E-06	16:66995927:G:A	0.00718015	0.07	0.02	7.05E-07	0	intronic	*CES3*
Hip_circumference	*TRAPPC4*	2	408182	5.83E-07	11:118890910:T:C	0.00017785	-0.52	0.10	6.94E-07	1	nonsynonymous	*TRAPPC4*
Waist_hip_ratio	*GPR151*	6	408144	3.00E-08	5:145895394:G:A	0.00834999	-0.05	0.01	8.08E-08	1	stopgain	*GPR151*
Height_standing	*PFDN2*	2	408034	1.41E-07	1:161071861:C:T	0.00044532	-0.25	0.05	8.46E-08	1	nonsynonymous	*PFDN2*
Body_mass_index	*FRMD5*	4	407605	1.92E-06	15:44167588:A:G	0.04908352	-0.03	0.01	1.09E-06	0	intronic	*FRMD5*
Body_mass_index	*IQSEC1*	12	407605	7.02E-07	3:12943013:T:G	0.00655782	0.08	0.02	6.91E-08	1	nonsynonymous	*IQSEC1*
Body_mass_index	*HCRTR2*	4	407605	1.50E-06	6:54854851:A:C	0.0145852	0.05	0.01	4.09E-07	0	intergenic	*FAM83B(dist=44954), HCRTR2(dist=116407)*
Weight	*FRMD5*	4	401786	6.24E-08	15:44147215:T:G	0.00644606	0.07	0.01	1.65E-07	0	intronic	*WDR76*
Weight	*HTRA1*	4	401786	2.00E-06	10:124193181:T:G	0.47045591	0.01	0.00	1.16E-07	0	intergenic	*PLEKHA1(dist=1310), ARMS2(dist=20998)*
Weight	*TRAPPC4*	2	401786	1.16E-06	11:118602242:C:T	0.00013775	-0.63	0.12	1.60E-07	0	intergenic	*TREH(dist=51861), DDX6(dist=16231)*
Whole_body_water_mass	*ZMYM6*	11	401782	1.75E-07	1:35298496:C:T	5.12E-05	1.18	0.23	1.59E-07	0	intergenic	*GJA4(dist=37148), SMIM12(dist=17467)*
Whole_body_water_mass	*FRMD5*	4	401782	2.98E-07	15:44564692:A:G	0.025069	-0.03	0.01	6.33E-08	0	intergenic	*FRMD5(dist=77200), CASC4(dist=16217)*
Whole_body_water_mass	*GLI3*	8	401782	1.09E-06	7:41650943:C:T	0.00041366	-0.21	0.04	1.22E-06	0	intergenic	*LINC01449(dist=477844), INHBA(dist=73769)*
Basal_metabolic_rate	*GPR151*	6	401771	1.97E-06	5:146212195:C:G	3.13E-05	1.09	0.23	1.74E-06	0	intronic	*PPP2R2B*

Basal_metabolic_rate	*ZMYM6*	11	401771	8.04E-07	1:35298496:C:T	5.12E-05	1.20	0.24	3.81E-07	0	intergenic	*GJA4(dist=37148), SMIM12(dist=17467)*
Whole_body_fat_free_mass	*ZMYM6*	11	401747	2.27E-07	1:35298496:C:T	5.12E-05	1.19	0.23	1.42E-07	0	intergenic	*GJA4(dist=37148), SMIM12(dist=17467)*
Whole_body_fat_free_mass	*FRMD5*	4	401747	5.88E-07	15:44564692:A:G	0.0250700 2	-0.03	0.01	7.16E-08	0	intergenic	*FRMD5(dist=77200), CASC4(dist=16217)*
Whole_body_fat_free_mass	*GLI3*	8	401747	1.94E-06	7:41650943:C:T	0.0004136 9	-0.21	0.04	1.34E-06	0	intergenic	*LINC01449(dist=477844), INHBA(dist=73769)*
Body_fat_percentage	*C10orf35*	2	401556	3.64E-07	10:71391560:A:G	0.0031652 3	0.10	0.02	2.26E-07	1	nonsynonymous	*C10orf35*
Days_per_week_moderate_phys_activity_10min	*RAD51AP1*	7	389204	3.87E-07	12:4657293:A:G	0.0065606 5	0.07	0.01	4.76E-07	1	nonsynonymous	*RAD51AP1*
Blood_pressure_diastolic_automated_mean	*ARID1B*	7	385365	1.08E-06	6:157525120:A:G	0.0012430 4	-0.19	0.04	9.01E-07	1	nonsynonymous	*ARID1B*
Pulse_rate_automated_mean	*DBH*	12	385365	1.74E-06	9:136149399:G:A	0.1869967 4	-0.02	0.00	3.46E-06	0	intronic	*ABO*
Blood_pressure_systolic_automated_mean	*TBX5*	4	385362	2.87E-07	12:114837349:C:A	0.0049278 3	-0.09	0.02	2.91E-07	1	nonsynonymous	*TBX5*

Table S3.6. Exome-wide significant genes with p-values ≤ $2.5x10^{-6}$ identified by SAIGE-GENE and remained significant after conditioning on the most significant variant, given that the most significant variant is a common variant with MAF > 1% or a less frequent non-coding variant that is not included in the gene-based tests for A. 53 quantitative traits B. 6 binary traits.

A.

						Most significant variant in the locus (+/- 500kb of the start and end positions of the gene)						
Phenotype	**Gene**	**Number of Variants**	**Sample Size**	**P-value (SKAT-O)**	**P-value (SKAT-O) conditional**	**chr:pos:ref:alt (GRCh37/hg19)**	**Allele frequency**	**BETA**	**SE**	**P-value**	**function**	**Gene**
Hip_circumference	GPR151	6	408,182	6.08E-07	6.13E-07	5:146212195:C:G	3.14E-05	1.636	0.336	1.12E-06	intronic	PPP2R2B
Hip_circumference	ANO1	4	408,182	5.98E-07	7.03E-07	11:69482091:C:A	0.646	0.015	0.003	1.77E-08	UTR3	ORAOV1 (NM_153451:c.*196G>T)
Waist_hip_ratio	SLC5A3	4	408,144	1.33E-07	6.23E-08	21:35593827:G:A	0.131	0.017	0.003	1.33E-09	intergenic	LINC00310(dist=31607),KCNE2(dist=142496)
Height_standing	C11orf57	2	408,034	5.71E-07	5.72E-07	11:111889209:T:C	2.71E-05	-1.410	0.226	4.42E-10	intronic	DIXDC1
Height_standing	CADM1	2	408,034	1.35E-07	1.36E-07	11:115841824:A:C	6.99E-05	0.870	0.141	6.91E-10	intergenic	LINC00900(dist=210906),LOC101929011(dist=668315)
Height_standing	PTH1R	3	408,034	1.40E-09	7.88E-11	3:46933960:C:G	0.581	-0.016	0.002	6.36E-18	intronic	PTH1R

Height_standing	SIX6	3	408,034	1.16E-06	3.30E-09	14:61072875:T:C	0.613	-0.022	0.002	3.07E-30	intergenic	SIX6(dist=94350),SALRNA1(dist=33059)
Height_standing	TBX3	2	408,034	4.59E-09	7.75E-08	12:115108136:T:C	0.257	0.016	0.002	1.75E-14	UTR3	TBX3
Height_standing	NOX4	4	408,034	4.09E-09	2.61E-07	11:89216425:T:A	0.796	0.015	0.002	9.65E-11	intronic	NOX4
Height_standing	S1PR5	5	408,034	1.44E-10	3.18E-10	19:10754905:G:T	0.664	-0.022	0.002	8.66E-29	UTR3	SLC44A2
Height_standing	MTMR11	3	408,034	1.94E-15	1.82E-20	1:149906413:T:C	0.409	0.035	0.002	7.59E-76	nonsynonymous	MTMR11
Height_standing	SERPINE2	3	408,034	2.80E-07	5.64E-07	2:225357325:T:C	0.004	0.224	0.016	5.68E-47	intronic	CUL3
Height_standing	ST3GAL4	5	408,034	5.37E-11	2.15E-10	11:125825224:G:T	0.582	0.015	0.002	1.81E-14	intergenic	DDX25
Height_standing	ABCB6	11	408,034	4.74E-08	1.38E-08	2:220004944:T:C	0.017	-0.139	0.008	2.50E-76	intronic	NHEJ1
Height_standing	TMEM150B	3	408,034	2.88E-07	1.36E-07	19:55993436:G:T	0.024	-0.101	0.006	2.83E-62	nonsynonymous	ZNF628
Height_standing	ZFAT	13	408,034	3.61E-32	7.36E-32	8:135598132:C:G	0.368	-0.023	0.002	2.53E-31	intronic	ZFAT
Height_standing	ELN	8	408,034	6.67E-08	2.92E-07	7:73474825:G:C	0.099	0.032	0.003	2.72E-24	nonsynonymous	ELN
Height_standing	AMOTL1	6	408,034	2.85E-07	2.87E-07	11:94547499:A:G	3.18E-03	0.099	0.017	3.21E-09	intronic	AMOTL1
Height_standing	COL11A1	8	408,034	2.91E-07	2.90E-07	1:103419238:A:G	0.700	-0.021	0.002	9.33E-24	intronic	COL11A1
Height_standing	PTCH1	15	408,034	2.65E-10	1.48E-09	9:98368761:T:C	0.258	0.044	0.002	7.44E-91	intergenic	PTCH1(dist=89514),LINC00476(dist=199609)
Height_standing	GRM4	2	408,034	7.23E-16	2.15E-10	6:34199092:C:T	0.911	-0.078	0.003	2.35E-125	intergenic	GRM4(dist=75693),HMGA1(dist=5485)
Height_standing	HAPLN3	6	408,034	1.25E-11	1.07E-11	15:89397827:A:G	0.030	-0.097	0.006	5.72E-69	intronic	ACAN
Height_standing	VASN	7	408,034	5.25E-07	5.16E-07	16:4019350:C:T	0.821	0.021	0.002	2.35E-18	intronic	ADCY9
Height_standing	PRSS56	3	408,034	1.68E-07	1.50E-07	2:232948266:A:G	0.916	0.048	0.003	1.64E-45	intronic	DIS3L2
Height_standing	TXLNA	3	408,034	5.66E-07	5.84E-07	1:32356815:C:T	0.061	0.029	0.004	7.70E-14	intergenic	SPOCD1(dist=75163),PTP4A2(dist=15207)
Height_standing	PHC3	7	408,034	3.24E-07	3.54E-07	3:169701455:A:G	3.29E-06	7.403	0.729	3.06E-24	intronic	SEC62
Height_standing	CLEC3A	3	408,034	9.08E-10	8.21E-10	16:78064682:G:A	1.87E-04	0.558	0.078	6.46E-13	nonsynonymous	CLEC3A
Height_standing	ESR1	9	408,034	7.28E-10	3.49E-07	6:152125444:T:C	0.661	0.020	0.002	5.95E-24	intronic	ESR1
Height_standing	SAMD4A	7	408,034	4.77E-09	1.15E-08	14:55238871:C:T	0.321	0.016	0.002	3.64E-16	intronic	SAMD4A
Height_standing	LRRC8A	2	408,034	2.87E-08	2.87E-08	9:131245464:G:A	2.42E-03	0.165	0.021	1.68E-15	ncRNA_intronic	MIR1268A
Height_standing	CPPED1	7	408,034	1.40E-07	1.32E-07	16:12399629:T:A	1.01E-03	0.211	0.035	1.82E-09	intronic	SNX29

Height_standing	FER	5	408,034	6.68E-12	1.12E-10	5:108171483:G:A	0.083	0.044	0.003	2.40E-38	synonymous	FER
Height_standing	LPP	8	408,034	2.11E-08	8.72E-08	3:187443314:G:A	0.071	0.036	0.004	1.48E-23	synonymous	BCL6
Height_standing	TRAPPC13	2	408,034	1.04E-10	1.15E-10	5:64766798:G:A	2.08E-03	-0.216	0.023	1.36E-21	nonsynonymous	ADAMTS6
Height_standing	CTU2	12	408,034	2.09E-08	1.92E-08	16:88782050:G:A	0.010	-0.086	0.010	9.03E-19	nonsynonymous	PIEZO1
Height_standing	PIEZO1	55	408,034	6.63E-08	2.21E-08	16:88782050:G:A	0.010	-0.086	0.010	9.03E-19	nonsynonymous	PIEZO1
Height_standing	PROP1	4	408,034	1.63E-07	2.89E-07	5:176980904:G:A	0.296	0.013	0.002	8.92E-10	intronic	FAM193B
Height_standing	FNDC3B	8	408,034	1.39E-12	6.82E-14	3:172188000:A:G	0.493	-0.025	0.002	4.54E-40	intergenic	GHSR(dist=21754), TNFSF10(dist=35298)
Height_standing	TARS2	3	408,034	4.22E-07	3.80E-09	1:149995265:G:A	0.398	0.035	0.002	7.13E-69	intergenic	OTUD7B(dist=12579), VPS45(dist=44085)
Height_standing	GYG1	5	408,034	1.78E-07	1.47E-07	3:148856741:A:C	0.064	-0.022	0.004	5.18E-09	intronic	HPS3
Height_standing	NUBP2	9	408,034	2.01E-22	2.51E-23	16:2160503:T:G	0.169	-0.030	0.003	9.96E-34	synonymous	PKD1
Height_standing	SPSB3	4	408,034	1.10E-22	1.19E-23	16:2160503:T:G	0.169	-0.030	0.003	9.96E-34	synonymous	PKD1
Height_standing	PKD1	43	408,034	6.40E-12	1.96E-08	16:2160503:T:G	0.169	-0.030	0.003	9.96E-34	synonymous	PKD1
Height_standing	C16orf70	2	408,034	7.45E-12	4.15E-11	16:67329937:T:C	0.074	-0.032	0.004	3.83E-19	intronic	KCTD19
Height_standing	ATAD5	8	408,034	4.74E-10	1.79E-07	17:29211667:G:A	0.270	-0.042	0.002	4.20E-87	intronic	ATAD5
Height_standing	CYR61	2	408,034	4.40E-09	7.38E-09	1:85997286:G:C	0.168	0.019	0.003	6.28E-13	intronic	DDAH1
Height_standing	STC2	4	408,034	1.10E-24	4.46E-25	5:172993684:C:G	0.631	-0.025	0.002	1.87E-37	intergenic	MIR8056(dist=219145), LOC285593(dist=12953)
Height_standing	PTK7	6	408,034	2.95E-07	6.24E-08	6:43002721:A:G	0.139	0.018	0.003	1.20E-11	intergenic	RRP36(dist=5384), CUL7(dist=2634)
Height_standing	ZCCHC6	7	408,034	1.42E-08	1.49E-07	9:89089476:C:T	0.503	0.016	0.002	3.03E-18	intergenic	ZCCHC6(dist=120074), GAS1(dist=469801)
Height_standing	NPR3	3	408,034	3.60E-25	4.14E-30	5:32711633:C:A	0.191	0.041	0.002	3.15E-61	UTR5	NPR3
Body_mass_index	TRAPPC4	2	407,605	2.87E-07	2.85E-07	11:118396331:A:T	0.023	-0.048	0.009	2.38E-08	ncRNA_intronic	LOC101929089
Weight	SCMH1	8	401,786	1.87E-07	3.69E-07	1:41570459:G:A	0.776	-0.022	0.003	1.42E-15	intronic	SCMH1
Weight	NUBP2	9	401,786	1.70E-07	7.64E-08	16:2160503:T:G	0.169	-0.027	0.003	4.02E-19	synonymous	PKD1
Weight	ZFAT	13	401,786	1.80E-11	3.38E-11	8:135612745:A:G	0.406	-0.018	0.002	5.67E-15	synonymous	ZFAT
Whole_body_water_mass	ZMYM6	11	401,782	1.75E-07	1.78E-07	1:35298496:C:T	5.12E-05	1.184	0.226	1.59E-07	intergenic	GJA4(dist=37148), SMIM12(dist=17467)
Whole_body_water_mass	FRMD5	4	401,782	2.98E-07	5.70E-07	15:44564692:A:G	0.025	-0.029	0.005	6.33E-08	intergenic	FRMD5(dist=77200), CASC4(dist=16217)

Whole_body_water_mass	LCOR	11	401,782	3.88E-07	5.08E-07	10:98657257:C:G	0.014	-0.045	0.007	6.72E-11	intronic	LCOR
Whole_body_water_mass	TMEM150B	3	401,782	2.96E-07	2.83E-07	19:55993436:G:T	0.024	-0.040	0.005	2.33E-14	nonsynonymous	ZNF628
Whole_body_water_mass	ESR1	9	401,782	1.15E-07	6.31E-07	6:152170247:G:A	0.455	0.014	0.002	2.55E-17	intronic	ESR1
Whole_body_water_mass	SCMH1	8	401,782	8.44E-18	1.58E-17	1:41570459:G:A	0.776	-0.022	0.002	2.47E-29	intronic	SCMH1
Whole_body_water_mass	NUBP2	9	401,782	1.52E-19	2.69E-20	16:2160503:T:G	0.169	-0.024	0.002	6.19E-29	synonymous	PKD1
Whole_body_water_mass	SPSB3	4	401,782	2.40E-17	3.72E-18	16:2160503:T:G	0.169	-0.024	0.002	6.19E-29	synonymous	PKD1
Whole_body_water_mass	ZFAT	13	401,782	1.83E-18	1.96E-17	8:135612745:A:G	0.406	-0.016	0.002	4.79E-22	synonymous	ZFAT
Basal_metabolic_rate	ZMYM6	11	401,771	8.04E-07	8.06E-07	1:35298496:C:T	5.12E-05	1.198	0.236	3.81E-07	intergenic	GJA4(dist=37148), SMIM12(dist=17467)
Basal_metabolic_rate	TMEM150B	3	401,771	1.10E-06	5.76E-07	19:55993436:G:T	0.024	-0.039	0.006	1.66E-12	nonsynonymous	ZNF628
Basal_metabolic_rate	FRMD5	4	401,771	1.35E-07	2.81E-07	15:44028047:C:T	0.024	-0.030	0.006	4.77E-08	downstream	CATSPER2P1(dist=99)
Basal_metabolic_rate	GRM4	2	401,771	9.53E-10	5.94E-07	6:34620153:T:A	0.139	0.035	0.002	1.86E-45	intronic	C6orf106
Basal_metabolic_rate	SCMH1	8	401,771	3.15E-15	5.90E-15	1:41570459:G:A	0.776	-0.022	0.002	8.01E-27	intronic	SCMH1
Basal_metabolic_rate	SPSB3	4	401,771	5.37E-15	9.79E-16	16:2160503:T:G	0.169	-0.025	0.002	1.21E-27	synonymous	PKD1
Basal_metabolic_rate	NUBP2	9	401,771	4.62E-17	9.62E-18	16:2160503:T:G	0.169	-0.025	0.002	1.21E-27	synonymous	PKD1
Basal_metabolic_rate	ZFAT	13	401,771	2.39E-17	1.46E-16	8:135612745:A:G	0.406	-0.016	0.002	3.07E-21	synonymous	ZFAT
Whole_body_fat_free_mass	ZMYM6	11	401,747	2.27E-07	2.04E-07	1:35298496:C:T	5.12E-05	1.187	0.226	1.42E-07	intergenic	GJA4(dist=37148), SMIM12(dist=17467)
Whole_body_fat_free_mass	TMEM150B	3	401,747	5.78E-07	2.88E-07	19:55993436:G:T	0.024	-0.041	0.005	1.40E-14	nonsynonymous	ZNF628
Whole_body_fat_free_mass	ESR1	9	401,747	7.69E-08	5.22E-07	6:152170247:G:A	0.455	0.014	0.002	8.43E-17	intronic	ESR1
Whole_body_fat_free_mass	SCMH1	8	401,747	2.90E-18	5.34E-18	1:41570459:G:A	0.776	-0.022	0.002	2.00E-29	intronic	SCMH1
Whole_body_fat_free_mass	SPSB3	4	401,747	2.51E-17	3.82E-18	16:2160503:T:G	0.169	-0.025	0.002	1.96E-29	synonymous	PKD1
Whole_body_fat_free_mass	NUBP2	9	401,747	1.70E-19	2.96E-20	16:2160503:T:G	0.169	-0.025	0.002	1.96E-29	synonymous	PKD1
Whole_body_fat_free_mass	ZFAT	13	401,747	5.30E-19	4.72E-18	8:135612745:A:G	0.406	-0.016	0.002	4.24E-22	synonymous	ZFAT
Impedance_of_whole_body	ADAMTS3	8	401,746	7.23E-12	2.12E-11	4:73519842:C:T	0.063	0.053	0.004	2.90E-39	intergenic	ADAMTS3(dist=85326), COX18(dist=400571)
Impedance_of_whole_body	SNED1	9	401,746	1.99E-06	1.36E-07	2:242050740:A:G	0.329	0.014	0.002	1.29E-11	intronic	PASK
Impedance_of_whole_body	NUBP2	9	401,746	3.02E-07	1.59E-07	16:2163962:A:G	0.098	0.025	0.003	1.92E-13	intronic	PKD1

Impedance_of_whole_body	FBN2	16	401,746	3.59E-07	9.05E-07	5:127367998:G:C	0.248	-0.032	0.002	1.53E-43	ncRNA_intronic	LINC01184
Impedance_of_whole_body	ZNF469	53	401,746	3.93E-08	1.73E-08	16:88321027:C:T	0.041	0.044	0.005	2.78E-18	intergenic	LINC02182(dist=92204), ZNF469(dist=172852)
Pulse_rate_automated_mean	ARHGEF40	7	385,365	7.02E-08	2.57E-10	14:21542766:A:G	0.169	0.051	0.003	3.30E-52	nonsynonymous	ARHGEF40
Pulse_rate_automated_mean	MYH6	14	385,365	3.61E-15	2.56E-13	14:23861811:A:G	0.370	0.072	0.003	1.04E-168	nonsynonymous	MYH6
Blood_pressure_diastolic_automated_mean	DBH	12	385,365	3.08E-14	5.20E-15	9:136522274:C:T	0.074	-0.041	0.005	4.91E-18	nonsynonymous	DBH
Blood_pressure_systolic_automated_mean	RRAS	2	385,362	1.33E-07	2.79E-07	19:49639399:C:G	0.155	-0.020	0.003	3.68E-10	intronic	PPFIA3
Blood_pressure_systolic_automated_mean	SLC9A3R2	5	385,362	2.36E-12	2.80E-13	16:2089006:G:A	0.159	-0.024	0.003	5.15E-14	UTR3	SLC9A3R2
Blood_pressure_systolic_automated_mean	DBH	12	385,362	2.85E-10	1.05E-10	9:136522274:C:T	0.074	-0.030	0.005	1.47E-11	nonsynonymous	DBH

B.

							Most significant variant in the locus (+/- 500kb of the start and end positions of the gene)						
Phenotype	**Phecode**	**Gene**	**Number of Variants**	**Number of cases**	**Number of controls**	**P-value (SKAT-O)**	**P-value (SKAT-O) conditional**	**chr:pos:ref:alt (GRCh37/hg19)**	**Allele frequency**	**Beta**	**SE**	**P-value**	**Function**	**Gene**
Cholelithiasis and cholecystitis	574	ABCB4	9	16225	391307	7.07E-07	1.43E-10	7:87105795:T:C	0.139	-0.148	0.017	3.29E-18	upstream	ABCB4(dist=776)
Cholelithiasis and cholecystitis	574	ABCG5	13	16225	391307	2.62E-13	5.35E-11	2:44069772:G:A	0.065	0.767	0.025	4.16E-201	intronic	ABCG8
Cholelithiasis and cholecystitis	574	ABCG8	12	16225	391307	4.68E-10	7.67E-11	2:44069772:G:A	0.065	0.767	0.025	4.16E-201	intronic	ABCG8
Diseases of hair and hair follicles	704	GORASP1	8	5344	402357	1.71E-09	3.72E-08	3:38659248:G:C	0.113	0.408	0.033	1.50E-35	intronic	SCN5A
Ankylosing spondylitis	715.2	IER3	4	620	365085	5.81E-10	1.36E-07	6:31210279:C:T	0.036	3.73	0.209	2.21E-71	intergenic	HCG27(dist=38534), HLA-C(dist=26247)
Ankylosing spondylitis	715.2	PSMB9	3	620	365085	7.07E-07	1.43E-10	6:32582577:A:C	0.340	0.61	0.062	1.10E-22	intergenic	HLA-DRB1(dist=24964), HLA-DQA1(dist=22606)

Table S3.7. Empirical type I error rates for SAIGE-GENE, SAIGE-GENE-GCadj (GC adjusted SAIGE-GENE), EmmaX-SKAT(Kang et al., 2010; M. C. Wu et al., 2011) and SMMAT(Chen et al., 2018). h^2: heritability.

		500 families and 5000 independent samples						1000 families and no independent samples					
		h^2=0.2			h^2=0.4			h^2=0.2			h^2=0.4		
	alpha	**burden**	**skat**	**skato**	**burden**	**skat**	**skato**	**burden**	**skat**	**skato**	**burden**	**skat**	**skato**
SAIGE-GENE	**0.05**	5.25E-02	5.62E-02	5.79E-02	5.49E-02	6.26E-02	6.32E-02	5.26E-02	5.62E-02	5.79E-02	5.48E-02	6.24E-02	6.30E-02
	0.0001	1.10E-04	1.30E-04	1.30E-04	1.40E-04	1.70E-04	1.80E-04	1.20E-04	1.30E-04	1.40E-04	1.40E-04	1.70E-04	1.80E-04
	2.50E-06	2.80E-06	4.20E-06	2.80E-06	3.50E-06	5.20E-06	3.30E-06	3.71E-06	5.22E-06	5.82E-06	3.81E-06	6.51E-06	5.61E-06
SAIGE-GENE -Gcadj	**0.05**	5.00E-02	5.00E-02	5.45E-02	5.01E-02	5.01E-02	5.10E-02	4.99E-02	4.99E-02	5.47E-02	4.99E-02	4.99E-02	5.10E-02
	0.0001	9.00E-05	9.00E-05	1.10E-04	1.00E-04	1.00E-04	8.00E-05	1.00E-04	1.00E-04	1.20E-04	1.00E-04	1.00E-04	8.00E-05
	2.50E-06	2.20E-06	2.20E-06	2.20E-06	1.90E-06	1.80E-06	1.30E-06	2.91E-06	2.91E-06	4.41E-06	2.51E-06	2.61E-06	2.51E-06
EmmaX-SKAT	**0.05**	5.16E-02	5.36E-02	5.58E-02	5.33E-02	5.75E-02	5.91E-02	5.18E-02	5.39E-02	5.60E-02	5.32E-02	5.76E-02	5.91E-02
	0.0001	1.10E-04	1.10E-04	1.20E-04	1.20E-04	1.30E-04	1.50E-04	1.10E-04	1.20E-04	1.30E-04	1.30E-04	1.40E-04	1.50E-04
	2.50E-06	2.36E-06	3.37E-06	2.13E-06	2.58E-06	3.03E-06	2.25E-06	3.71E-06	4.61E-06	5.40E-06	3.49E-06	5.06E-06	4.95E-06
SMMAT	**0.05**	5.17E-02	5.36E-02	5.41E-02	5.33E-02	5.74E-02	5.73E-02	5.18E-02	5.39E-02	5.43E-02	5.32E-02	5.76E-02	5.74E-02
	0.0001	1.10E-04	1.10E-04	1.40E-04	1.20E-04	1.30E-04	1.70E-04	1.10E-04	1.20E-04	1.50E-04	1.30E-04	1.40E-04	1.70E-04
	2.50E-06	2.50E-06	3.80E-06	2.90E-06	2.80E-06	3.40E-06	3.30E-06	3.47E-06	4.42E-06	6.00E-06	3.26E-06	4.63E-06	5.47E-06

Table S3.8. Empirical type I error rates for SAIGE-GENE with the larger sample size of 1,000 families and 10,000 independent samples (total sample size N = 20,000). The heritability $h^2 = 0.2$.

	alpha	Burden	SKAT	SKAT-O
SAIGE-GENE	**0.05**	5.28E-02	5.68E-02	5.84E-02
	0.0001	1.00E-04	1.00E-04	1.00E-04
	2.50E-06	3.44E-06	4.33E-06	3.00E-06

Table S3.9. Empirical type I error rates for SAIGE-GENE, EmmaX-SKAT(Kang et al., 2010; M. C. Wu et al., 2011) and SMMAT(Chen et al., 2018) for skewed distributed phenotypes with and without inverse normal transformation for 500 families and 5,000 independent samples (total sample size N = 10,000). The heritability $h^2 = 0.2$.

		with inverse normal transformation			without inverse normal transformation		
	alpha	**Burden**	**SKAT**	**SKAT-O**	**Burden**	**SKAT**	**SKAT-O**
SAIGE-GENE	**0.05**	5.19E-02	5.61E-02	5.75E-02	5.13E-02	5.85E-02	5.87E-02
	0.0001	1.25E-04	1.75E-04	1.81E-04	1.52E-04	3.50E-04	3.27E-04
	2.50E-06	4.44E-06	8.27E-06	6.65E-06	5.68E-06	2.86E-05	2.19E-05
EmmaX-SKAT	**0.05**	5.10E-02	5.37E-02	5.55E-02	5.09E-02	5.75E-02	5.79E-02
	0.0001	1.18E-04	1.55E-04	1.61E-04	1.45E-04	3.48E-04	3.30E-04
	2.50E-06	4.82E-06	6.55E-06	6.55E-06	6.92E-06	2.42E-05	2.29E-05
SMMAT	**0.05**	5.14E-02	5.41E-02	5.45E-02	5.09E-02	5.72E-02	5.59E-02
	0.0001	1.19E-04	1.62E-04	1.83E-04	1.42E-04	3.40E-04	3.40E-04
	2.50E-06	5.42E-06	8.34E-06	9.17E-06	6.79E-06	2.59E-05	2.59E-05

Table S3.10. Empirical type I error rates for SAIGE-GENE and SMMAT(Chen et al., 2018) for skewed distributed phenotypes with the three-step phenotype transformation procedure for 500 families and 5,000 independent samples (total sample size N = 10,000). The heritability $h^2 = 0.2$.

	alpha	Burden	SKAT	SKAT-O
SAIGE-GENE	**0.05**	4.71E-02	4.42E-02	4.77E-02
	0.0001	8.53E-05	7.94E-05	9.32E-05
	2.50E-06	1.79E-06	2.53E-06	2.74E-06
SMMAT	**0.05**	5.03E-02	5.06E-02	5.16E-02
	0.0001	1.03E-04	1.05E-04	1.29E-04
	2.50E-06	1.71E-06	2.32E-06	3.54E-06

Table S3.11. Empirical type I error rates for SAIGE-GENE in the presence of population stratification. Phenotypes were simulated based on real genotypes from the UK Biobank for randomly selected 5,000 samples with white British ancestry and 5,000 samples with European ancestry but not white British. The heritability $h^2 = 0.2$.

alpha	Burden	SKAT	SKAT-O
0.05	5.05E-02	5.09E-02	5.33E-02
0.0001	1.02E-04	1.06E-04	1.33E-04
2.50E-06	2.42E-06	2.42E-06	3.85E-06

Table S3.12. Empirical type I error rates for SAIGE-GENE in the presence of non-negligible cryptic relatedness. Phenotypes were simulated based on real genotypes from the UK Biobank for randomly selected 10,000 sample with White British ancestry (5,000 are related with up to 3rd degree and 5,000 are unrelated). The heritability $h^2 = 0.2$.

alpha	Burden	SKAT	SKAT-O
0.05	5.07E-02	5.11E-02	5.35E-02
0.0001	1.06E-04	1.07E-04	1.45E-04
2.50E-06	3.04E-06	2.88E-06	3.80E-06

Table S3.13. Empirical Type I error rates of SAIGE-GENE for binary traits with five different prevalence. Phenotypes were simulated for 500 families and 5,000 independent samples (total sample size N = 10,000). The liability scale heritability h^2_{latent} =0.23. A. Unadjusted SAIGE-GENE; B. SAIGE-GENE with the robust adjustment to account for unbalanced case-control ratios. SKAT-O-GCadj and SKAT-GCadj are GC-adjusted SKAT-O and SKAT.

A. Unadjusted SAIGE-GENE without applying the robust adjustment.

	Alpha	Prev=0.01	Prev=0.05	Prev=0.1	Prev=0.2	Prev=0.5
SKAT-O	**0.05**	7.27E-02	5.55E-02	5.38E-02	5.31E-02	5.32E-02
	0.0001	2.77E-03	4.57E-04	2.32E-04	1.43E-04	1.04E-04
	2.50E-06	5.38E-04	4.04E-05	1.49E-05	5.76E-06	1.70E-06
SKAT	**0.05**	8.67E-02	5.59E-02	5.21E-02	5.05E-02	5.03E-02
	0.0001	3.10E-03	4.74E-04	2.33E-04	1.30E-04	8.96E-05
	2.50E-06	6.34E-04	4.64E-05	1.58E-05	6.69E-06	1.70E-06
Burden	**0.05**	4.57E-02	4.94E-02	4.99E-02	5.03E-02	5.06E-02
	0.0001	5.69E-04	1.85E-04	1.27E-04	1.08E-04	1.01E-04
	2.50E-06	7.09E-05	1.16E-05	4.80E-06	3.60E-06	2.90E-06

B. SIAGE-GENE with the robust adjustment.

	Alpha	Prev=0.01	Prev=0.05	Prev=0.1	Prev=0.2	Prev=0.5
SKAT-O	**0.05**	5.52E-02	5.06E-02	4.77E-02	4.39E-02	4.12E-02
	0.0001	3.14E-04	1.58E-04	1.16E-04	8.10E-05	5.51E-05
	2.50E-06	7.30E-06	4.51E-06	3.80E-06	1.54E-06	1.10E-06
SKAT-O-GCadj	**0.05**	5.00E-02	5.00E-02	4.77E-02	4.39E-02	4.12E-02
	0.0001	2.08E-04	1.51E-04	1.16E-04	8.10E-05	5.51E-05
	2.50E-06	3.90E-06	4.51E-06	3.80E-06	1.54E-06	1.10E-06
SKAT	**0.05**	6.59E-02	5.01E-02	4.49E-02	3.97E-02	3.63E-02
	0.0001	3.40E-04	1.53E-04	1.13E-04	6.76E-05	3.94E-05
	2.50E-06	8.20E-06	4.11E-06	2.90E-06	1.65E-06	5.01E-07
SKAT-GCadj	**0.05**	5.00E-02	5.00E-02	4.49E-02	3.97E-02	3.63E-02
	0.0001	1.01E-04	1.53E-04	1.13E-04	6.76E-05	3.94E-05
	2.50E-06	9.00E-07	4.11E-06	2.90E-06	1.65E-06	5.01E-07
Burden	**0.05**	3.95E-02	4.67E-02	4.66E-02	4.54E-02	4.42E-02
	0.0001	7.82E-05	8.58E-05	7.65E-05	7.19E-05	6.41E-05
	2.50E-06	1.80E-06	2.91E-06	2.00E-06	1.54E-06	2.10E-06

Table S3.14. Empirical type I error rates for SAIGE-GENE in the simulation study with case-control sampling from an underlying large cohort. The liability scale heritability h^2_{latent} =0.23.

		Case:Control=1:1			Case:Control=1:9		
	alpha	**Burden**	**SKAT**	**SKAT-O**	**Burden**	**SKAT**	**SKAT-O**
SAIGE-GENE	**0.05**	4.97E-02	4.86E-02	5.17E-02	4.87E-02	4.96E-02	5.16E-02
	0.0001	9.21E-05	9.05E-05	1.01E-04	9.07E-05	1.18E-04	1.27E-04
	2.50E-06	2.68E-06	1.56E-06	2.23E-06	2.37E-06	3.85E-06	2.07E-06

Table S3.15. Empirical power for SAIGE-GENE and EmmaX-SKAT with two different percentages of causal variants (top vs bottom panels) and two different ratios of positive and negative effect directions (left vs right). β: effect size. h^2: heritability.

h^2=0.4	Proportion of causal variants= 0.4					
	β −/+=0.8/0.2			β −/+=1/0		
	Burden	**SKAT**	**SKAT-O**	**Burden**	**SKAT**	**SKAT-O**
EmmaX-SKAT	54.80%	80.20%	84.20%	89.10%	82.1%	92.00%
SAIGE-GENE	54.60%	81.30%	84.00%	89.00%	83.20%	91.70%
	Proportion of causal variants= 0.1					
	β −/+=0.8/0.2			β −/+=1/0		
	Burden	**SKAT**	**SKAT-O**	**Burden**	**SKAT**	**SKAT-O**
EmmaX-SKAT	35.3%	66.30%	66.60%	51.70%	65.6%	66.10%
SAIGE-GENE	35.20%	67.60%	66.50%	52.00%	67%	66.40%

Table S3.16. Empirical power for the SKAT-O test in SAIGE-GENE for binary phenotypes in cohort studies and case-control sampling studies. 40% of variants were simulated to be causal. 80% of causal variants were risk-increasing variants, while 20% were risk-decreasing variants. The liability scale heritability h^2_{latent} =0.23. Note that the effect sizes in cohort study and case-control sampling were different, so the power in these two designs is not directly comparable.

	Cohort study		Case-control sampling	
case:control	**Unadjusted SKAT-O**	**Robust SKAT-O**	**Unadjusted SKAT-O**	**Robust SKAT-O**
1:1	56.00%	58.00%	60.00%	61.00%
1:9	44.00%	56.00%	21.00%	23.00%
1:19	27.00%	37.00%	8.00%	9.00%
1:99	5.00%	8.00%		

CHAPTER IV

The Construction of Multi-ethnic Polygenic Risk Score Using Transfer Learning

Abstract

Methods for constructing polygenic risk scores (PRS), including pruning and thresholding (PT), lassosum (lsum) and PRS-CS, have been extensively investigated in recent years. As most existing GWAS were conducted in European or East Asian individuals, the existing PRS models have limited transferability to minority populations such as Africans and South Asians. Although recent studies have developed multi-ethnic PRS models that linearly combine multiple PRS trained with different ancestry GWAS, they remain under-powered.

Here we propose a novel multi-ethnic PRS using transfer learning techniques, borrowed from the machine learning literature. Our approach, TL-PRS, fine-tunes the potentially biased model trained with GWAS summary statistics from the majority/source ancestry to the target dataset of the minority ancestry. TL-PRS can use any existing PRS methods (such as lsum and PRS-CS) as a baseline method for fine-tuning. Using the potentially biased baseline parameter estimates as initial values, TL-PRS iterates the gradient descent algorithm to adapt the parameters for the target ancestry group. In the presence of multiple GWAS summary sources from different ancestries,

TL-PRS combines fine-tuned PRS using linear combination.

Through simulation studies, we show that TL-PRS improved the performance of PRS across a wide range of genetic architectures and cross-population genetic correlations. TL-PRS was most effective when genetic correlations between populations and samples used for calculating summary statistics were small and the genetic architecture was less polygenic. For example, when genetic correlation was 0.4, TL-PRS-lsum attained on average a 228% and 36% relative improvement in prediction accuracy compared to lsum when the causal variants were 0.1% and 1% respectively while TL-PRS-CS attained on average a 26% and 20% relative improvement compared to PRS-CS.

In our application using 8,168 African samples from the UK Biobank data, TL-PRS substantially improved the prediction accuracy of the six quantitative and two dichotomous traits. Compared to lsum, TL-PRS-lsum attained a 21% and 26% average relative improvement in prediction accuracy when using Biobank Japan and UK Biobank white British GWAS summary statistics, respectively; the average relative improvement of TL-PRS-CS over PRS-CS was 40% and 9%, respectively. When combining summary statistics from Biobank Japan and UK Biobank, TL-PRS-lsum, and TL-PRS-CS outperformed the linear combination of PT and PRS-CSx. By improving the polygenic risk prediction in non-European individuals, our approach will increase the usefulness of PRS and help reduce potential health inequities.

4.1 Introduction

Genetic risk prediction is one of the important and widely investigated topics in genetic epidemiology because it can help us better understand the genetic architecture of complex traits and potentially aid the clinical decision-making(Lewis & Vassos, 2020; Sugrue & Desikan, 2019; Torkamani et al., 2018). Many polygenic risk prediction methods have been developed and applied to complex traits, such as pruning and thresholding (PT)(Vilhjálmsson et al., 2015), lassosum (lsum)(Mak et al., 2017), PRS-CS(Ge et al., 2019) and LDpred(Vilhjálmsson et al., 2015).

In the prediction of polygenic risk in European-ancestry populations, these methods perform well and help to identify high risk groups because of the large sample size of Europeans in the training data (Duncan et al., 2019; Martin et al., 2019; Torkamani et al., 2018; Vilhjálmsson et al., 2015). However, due to insufficient non-European training data, these methods are inadequate for the polygenic risk prediction in non-European populations (Ge et al., 2019; Mak et al., 2017). Directly using PRS models trained by European samples for the prediction of non-European groups may reduce the prediction accuracy for genetic differences across ancestry groups (Duncan et al., 2019; Vilhjálmsson et al., 2015).

To tackle this issue, Márquez-Luna et al. proposed a multi-ethnic polygenic risk scores (PRS) model by linearly combining the PRS of two training populations(Márquez-Luna et al., 2017). They attained more than 70% relative improvement in prediction accuracy for type 2 diabetes in both Latino and South Asian cohorts compared to prediction methods using training data from a single population. However, this method implicitly assumes that the effect sizes from two populations are linearly combined with a constant weight for all SNPs and this linear assumption may not hold in all situations.

To address this problem, we propose a novel multi-ethnic PRS using transfer learning techniques, borrowed from the machine learning literature. Transfer learning is a tool in machine learning that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem.(West et al., 2007) The procedure of transfer learning is an optimization that allows rapid progress or improved performance when modeling the second task.(Pan & Yang, 2009; Torrey & Shavlik, 2010) From the practical viewpoint, reusing or transferring information from previously learned tasks for the learning of new tasks has the potential to significantly improve the prediction performance as well as model efficiency(West et al., 2007).

Our approach, Transfer Learning PRS or TL-PRS, fine-tunes the potentially biased model trained with GWAS summary statistics from majority ancestry to the target dataset of minority ancestry. TL-PRS can use any existing PRS methods (such as lsum and PRS-CS) as a baseline method for fine-tuning. Using the potentially biased baseline parameter estimates as initial values, TL-PRS iterates the gradient descent algorithm to adapt the parameters for the target ancestry group. In addition, in the presence of multiple GWAS summary statistics from different ancestries, TL-PRS combines fine-tuned PRS using linear combination. The implementation is also scalable for large data analysis.

In the simulations, TL-PRS outperformed the existing PRS models in a wide range of genetic architectures and cross-population genetic correlations. In the application, we use 10,285 South Asian samples and 8,168 African samples from the UK Biobank data. The prediction accuracy of many traits was improved by TL-PRS. By improving the polygenic risk prediction in non-European individuals, our approach will improve the performance of PRS and help reduce potential health inequities.

4.2 Methods

4.2.1 Polygenic risk score using GWAS summary statistics from a single ancestry

Assuming that summary statistics (i.e. the estimated effect size $\hat{\beta}_j$) are available, PRS is constructed as the summation of the estimated effects across all single-nucleotide polymorphisms (SNPs) on a given phenotype. Given a typical individual i, PRS can be defined as

$$PRS_i = \sum_{j=1}^{M} \hat{\beta}_j G_{ij}.$$

Although PRS_i aggregate the effects of all genetic variants to phenotypes, it has two major limitations: (1) The LD correlation among SNPs is ignored and this formula implicitly assumes that all SNPs are independent. (2) Some SNPs may not contribute to the phenotype; therefore, including them in the model reduces the prediction accuracy.

There are several well-known methods to overcome the limitations, such as pruning and thresholding (PT), lassosum (lsum), and PRS-CS. PT computes the PRS on a subset of SNPs based on LD-pruning and P-value thresholding. It has two tuning parameters: a pairwise threshold R^2_{LD} and a p-value threshold P_T. Lsum re-estimates the effect sizes using elastic net on GWAS summary statistics. The hyperparameters include the coefficients of L1 and L2 penalty. PRS-CS is a Bayesian polygenic prediction approach that uses a continuous shrinkage prior to derive posterior SNP effect sizes from summary statistics. The global shrinkage parameter in PRS-CS is a hyper-parameter that models the overall sparseness of the genetic architecture. Overall, PT and lsum are computationally efficient while PRS-CS requires more computational time due to the Bayesian

framework. In terms of prediction accuracy, lsum and PRS-CS generally outperform than PT in most circumstances.

4.2.2 Transfer learning (TL-PRS) using GWAS summary statistics from a single ancestry

Suppose that we have trained a PRS model using summary statistics from source population A, this model could be considered as prior knowledge to predict the genetic effects in the target population B. However, due to different LD patterns and possible effect size heterogeneity across ancestries, effect size estimation from population A can be viewed as biased estimators of effect sizes in population B. In order to reduce the bias and achieve better prediction performance, we borrow the idea of transfer learning and attempt to utilize a small fraction of the target sample to combine information from the biased model and the target sample data.

Specifically, for the target population, we have the following model:

$$Y = \sum G_j \beta_j + C\gamma + \varepsilon = \sum G_j (\beta_j^u + \tau_j) + C\gamma + \varepsilon,$$

where β_j is the true effect size of the target population, assumed to be unknown; β_j^u is given by the trained model, assumed to be biased; τ_j is the bias term between β_j^u and β_j; C is the covariate matrix including the intercept; and γ is a vector of covariate coefficients. Since this problem can be converted into a convex optimization, we can perform gradient descent algorithms on β_j with the initial value β_j^u using the following formula:

$$\beta_j{}^{(i+1)} = \frac{G_j(Y - C\gamma) - G_j^T G\beta^{(i)} + G_j^T G_j \beta_j{}^{(i)}}{G_j^T G_j} = \beta_j{}^{(i)} + \alpha \frac{G_j(Y - C\gamma) - G_j^T G\beta^{(i)}}{G_j^T G_j},$$

where $G_j Y$ and $G_j^T G$ can be pre-calculated using the training data and α is the learning rate. In addition, early stopping of iteration is required to avoid overfitting.

Both the learning rate α and the number of iterations n_{stop} can be tuned based on the validation dataset in terms of the best prediction accuracy. In order to reduce computation cost, we suggest choosing α from a small grid of values $min\left(\frac{1,10,100,1000}{N(SNPs)}, 1\right)$.

Additionally, any PRS method can be used for initial value β_j^u for transfer learning. In this paper, we used lsum and PRS-CS trained models as the baseline methods, which are referred as TL-PRS-lsum and TL-PRS-CS, respectively. TL-PRS method can also be applied to other training models, such as LDpred(Vilhjálmsson et al., 2015).

4.2.3 Combining multiple GWAS summary statistics from different ancestries

Suppose we have constructed the PRS from two different GWAS summary statistics PRS_1 and PRS_2, then the multi-ethnic PRS can be built as

$$PRS = \pi PRS_1 + (1 - \pi) PRS_2,$$

where π is a tuning parameter with range [0,1] and can be decided using the cross-validation method.(Márquez-Luna et al., 2017) This idea was first proposed by Márquez-Luna et al in 2017 using PT to construct a single-ancestry PRS, which was referred as PT-multi. We can extend this approach to other PRS construction methods, such as lsum and PRS-CS, which are lsum-multi, PRS-CSx(Huang et al., 2021). Similarly, the linear combination can also be applied to TL-PRS models. For example, if we have TL-PRS-lsum models from two populations, we might linearly combine them and refer to this method as MTL-PRS-lsum. MTL-PRS-CS can also be constructed in the same way.

Beyond the combination of two populations, we can further extend this idea to three or more different ancestries. Suppose that we have PRS from three different populations PRS_1, PRS_2 and PRS_3, then the multi-ethnic PRS can be built as

$$PRS = \pi_1 PRS_1 + \pi_2 PRS_2 + (1 - \pi_1 - \pi_2) PRS_3, \text{ where } \pi_1, \pi_2 \geq 0 \text{ and } \pi_1 + \pi_2 \leq 1.$$

4.2.4 Simulations using South Asian samples in UK Biobank

We simulated quantitative phenotypes using real genotypes from the South Asian samples in the UK Biobank. The proportion of causal markers was fixed as 1% and 0.1%, the SNP-heritability ${h_g}^2$ was fixed at 0.5. The normalized effect sizes b_i were sampled from a normal distribution with mean 0 and variance equal to ${h_g}^2$ divided by the number of causal markers. The per-allele effect sizes could be calculated by $\beta_i = \frac{b_i}{\sqrt{2p_i(1-p_i)}}$, where p_i is the minor allele frequency of i-th SNP. We simulated phenotypes as

$$Y_j = \sum_{i=1}^{M} \beta_i g_{ij} + \epsilon_j, \text{ where } \epsilon_j \sim N(0, 1 - {h_g}^2).$$

M is the number of SNPs and only HapMap3 variants (Consortium, 2010) are included in the simulation.

The GWAS summary statistics based on 10,000 South Asians and 100,000 Europeans were generated respectively based on the formula $\hat{\beta}_i \sim N(\hat{R}\beta_i, \hat{R}/n)$, where n is the sample size and $\hat{R}$ is the estimated correlation matrix of the LD block region using South Asians and Europeans in 1000 Genomes Project. We assumed that causal variants could be shared across all populations (Europeans and South Asians), but varying effect sizes were allowed and sampled from a

multivariate normal distribution with a genetic correlation of 0.4, 0.7, or 1.0. The simulation of the phenotype was repeated 20 times.

The target South Asian data set was randomly split into training, validation and testing datasets. In the training and validation dataset, we applied single-source prediction methods (PT, lsum, PRS-CS) as well as TL-PRS methods (TL-PRS-lsum, TL-PRS-CS) to GWAS summary statistics generated by 10K South Asians and 100K Europeans samples and evaluated their predictive performance measured by R^2 between the simulated and predicted phenotypes in the testing set. The multi-source prediction methods (PT-mutli, lsum-multi, MTL-PRS-lsum, PRS-CSx, MTL-PRS-CS) were then utilized to compare the cross-population polygenic prediction.

4.2.5 Analysis of South Asian, African, and non-British White samples in UK Biobank

We constructed PRSs for following target samples in UK Biobank: South Asian (SAS), African (AFR) and Non-British White (NBW). In each target sample, we used the software KING to exclude one individual in each related pair (up to second-degree relatives). We then built the polygenic prediction models on the following eight traits: high-density lipoproteins (HDL), low-density lipoproteins (LDL), body mass index (BMI), triglycerides (TG), systolic blood pressure (SBP), diastolic blood pressure (DBP), coronary artery disease (CAD), and Type II diabetes (T2D). The first six traits were quantitative and the last two traits were dichotomous.

Summary statistics of GWAS analyses on White British in UK Biobank (UKBB) and Japanese in Biobank Japan (BBJ) were downloaded from UKBB (https://pheweb.org/UKB-Neale/) and BBJ PheWeb (http://jenger.riken.jp/en/result). We restricted our analysis to common variants

(MAF>=0.01) presented in summary data and target genotype files after removing A/T and C/G SNPs to eliminate potential strand ambiguity (Márquez-Luna et al., 2017).

We applied single-source prediction methods (PT, lsum, TL-PRS-lsum, PRS-CS, TL-PRS-CS) to UKBB and BBJ summary statistics, and used multi-source prediction methods (PT-multi, lsum-multi, MTL-PRS-lsum, PRS-CSx, MTL-PRS-CS) to combine UKBB and BBJ GWAS results. The prediction accuracy was assessed in the testing dataset of each target population separately, adjusting for age, sex and the top four principal components (PCs). We used R^2 as the prediction accuracy metric for quantitative traits and Nagelkerke R^2 for dichotomous traits. For each population, the target samples were randomly split into a training dataset (for model fitting), a validation dataset (for hyper-parameter tuning) and a testing dataset (for the evaluation of predictive performance).

4.3 Results

4.3.1 Overview of TL-PRS

We first build PRS models using existing methods, such as lsum and PRS-CS. These models provide biased effect sizes of causal SNPs, which are used as the initial values of TL-PRS. The hyperparameters in TL-PRS include the learning rate and the number of iterations. Given TL-PRS models from different GWAS summary sources, we can integrate them by learning an optimal linear combination to produce the final PRS (Figure 4.1).

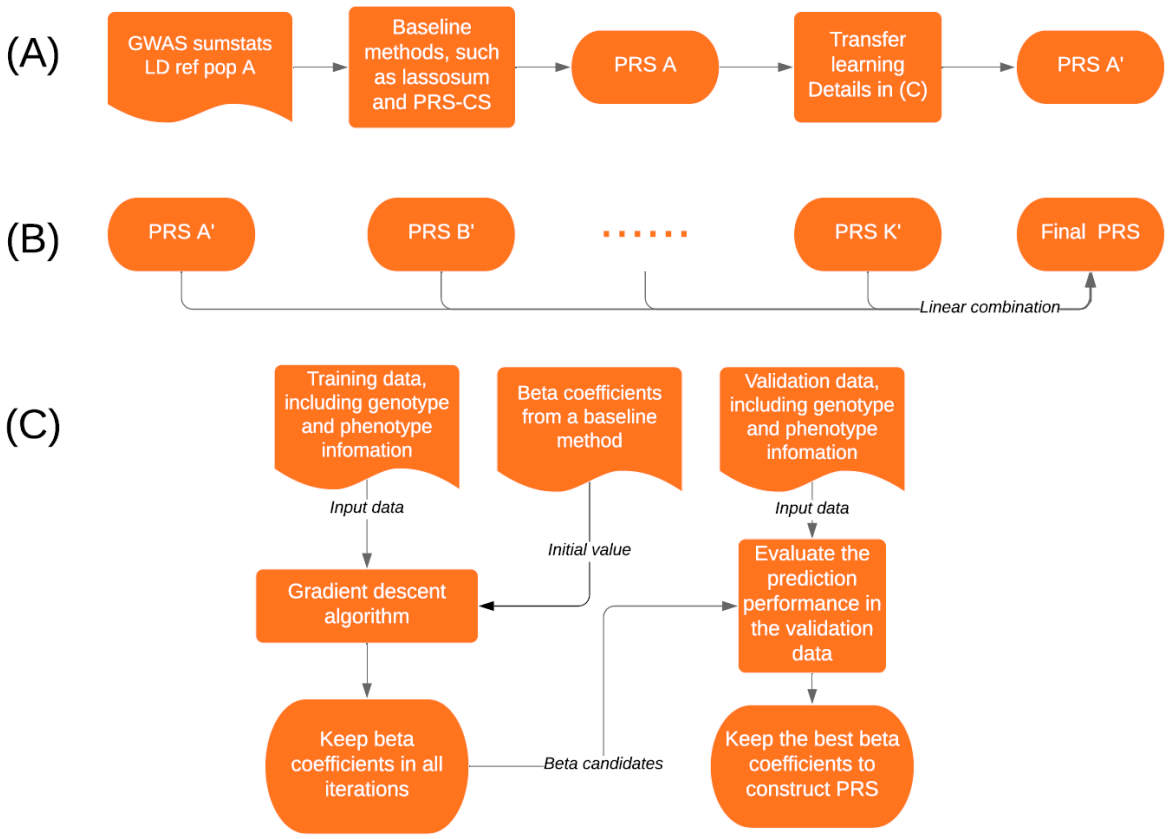


Figure 4.1: Overview of transfer learning on PRS methods. LD ref: LD reference panel.
(A) The general procedure how to construct PRS using transfer learning;
(B) The procedure how to combine multiple PRS into the final PRS;
(C) The detailed procedure of transfer learning.

Figure 4.2 shows the relative accuracy (R^2_{TL}/R^2_{Base}) of TL-PRS as a function of iterations. The relative accuracy in the training dataset kept increasing with respect to the iterations, which caused the overfitting. However, the fifth iteration reached the maximum relative accuracy in the validation sets of both simulation and real data analysis, which suggested that the fifth iteration was the optimal point to stop in these two examples. A similar strategy can be applied to choose the learning rate.

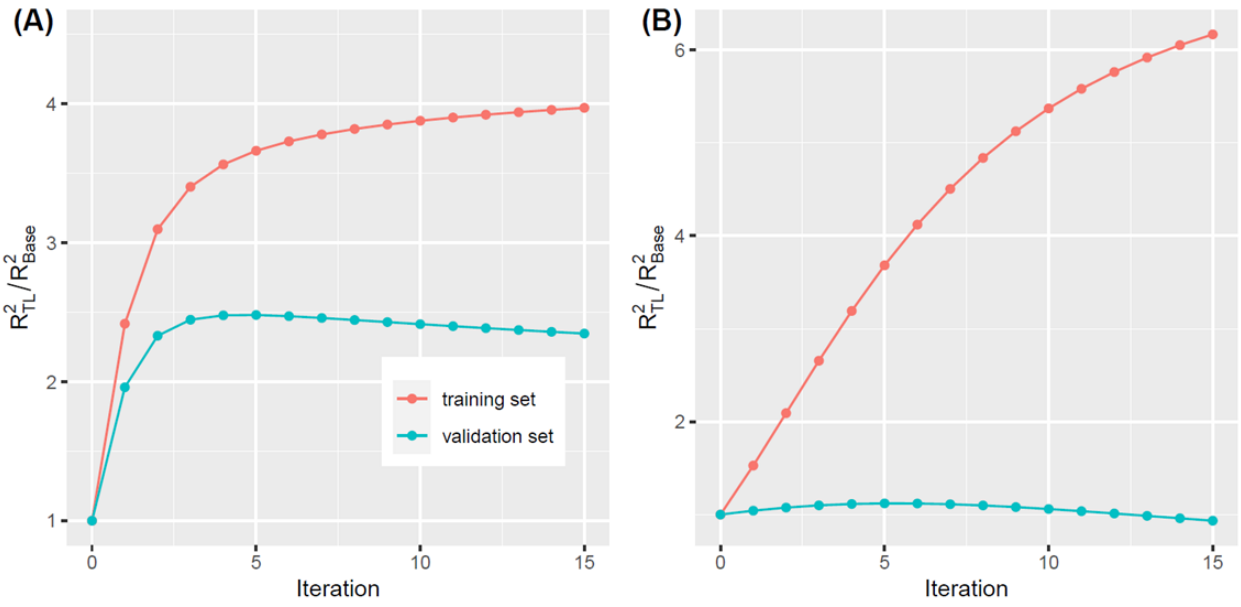


Figure 4.2. Relative accuracy of transfer learning method as a function of iterations.
(A) The simulation setting where the causal markers were 0.1%, genetic correlation was 0.4 and European summary statistics were used.
(B) The real data analysis of HDL in a South Asian cohort from UK Biobank, where UKBB summary statistics were used.

4.3.2 Simulations using South Asian samples in UK Biobank

In the simulation, different scenarios were considered by randomly selecting 0.1% or 1% variants across the genome as causal variants, which explained 50% of the phenotypic variance in total. Additionally, causal variants were assumed to be the same across populations, but different effect sizes were simulated from a multivariate normal distribution using the cross-population genetic correlation 0.4, 0.7 and 1. We generated 20 datasets in each scenario to evaluate the predictive performance of different PRS construction methods.

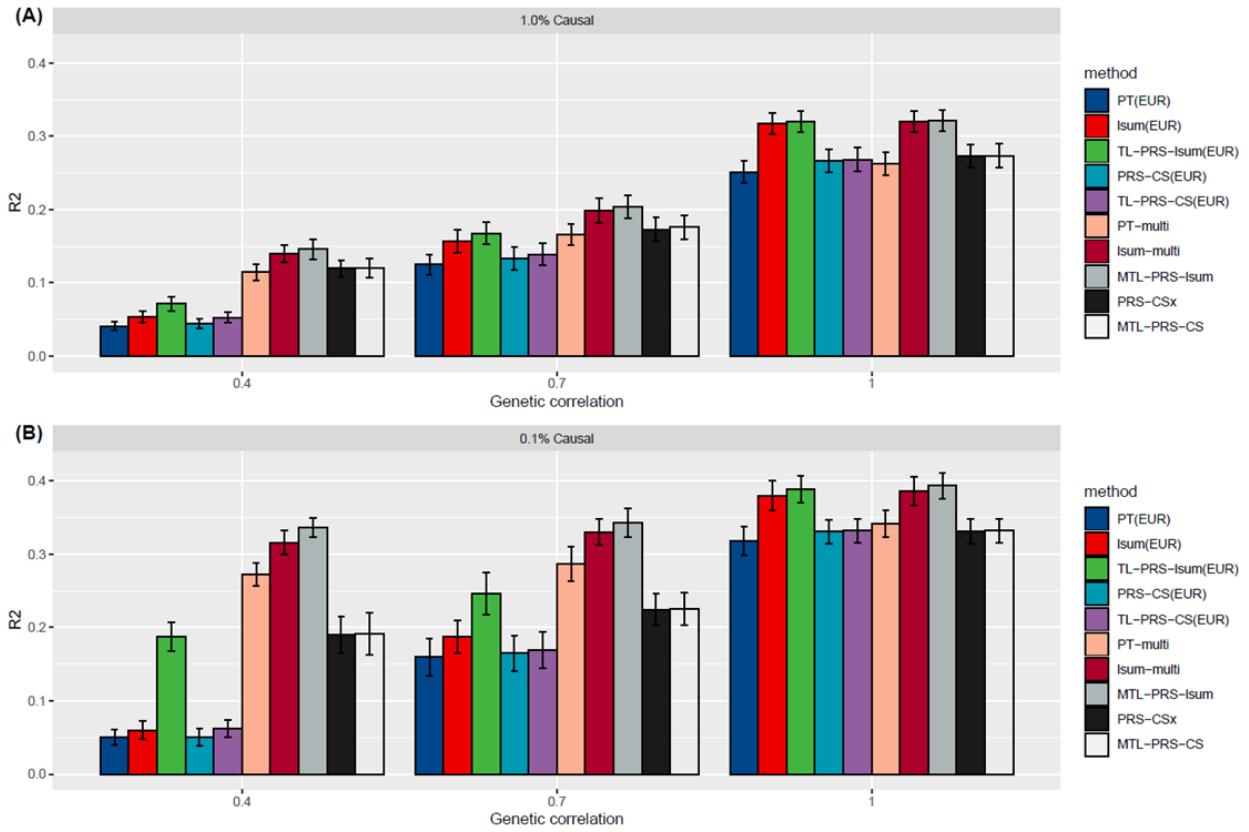


Figure 4.3. Prediction accuracy of single-source and multi-source polygenic prediction methods in simulations. Different genetic designs (0.1% and 1% causal variants) were simulated as well as cross-population genetic correlations (0.4, 0.7 and 1.0). Heritability was fixed at 50%. Prediction accuracy was measured by the squared correlation (R^2) between the simulated and predicted phenotypes in the testing dataset, averaged across 20 simulation replicates. Error bar indicates the standard deviation of R^2 across simulation replicates.

The prediction accuracy of single-source and multi-source polygenic prediction methods in the simulations can be found in Figure 4.3. For a fixed heritability 0.5, the predictive performance of all ten PRS methods increased when the genetic architecture became less polygenic (1.0% vs 0.1% causal). Although the causal variants were identical between populations, all ten PRS methods decreased prediction accuracy when the genetic effects were less correlated among populations. This is also the situation where TL-PRS could further improve the prediction accuracy. For example, when genetic correlation was 0.4, TL-PRS-lsum improved a 228% and 36% average

prediction accuracy compared to lsum when the causal variants were 0.1% and 1%, respectively (Figure 4.4). The relative improvement of TL-PRS-CS over PRS-CS was 26% and 20% on average. However, when genetic correlation was 1.0, lsum and PRS-CS are sufficient for prediction in target population due to the unbiased effect sizes. TL-PRS-lsum and TL-PRS-CS could attain limited relative improvement of the prediction accuracy in this situation. In general, TL-PRS worked better when the genetic correlation was smaller and the causal variants were sparser.

We further assessed whether multi-source prediction methods (PT-multi, lsum-multi, MTL-PRS-lsum, PRS-CSx, MTL-PRS-CS) could improve cross-population polygenic prediction. Specifically, we combined PRS models from European summary source (N=100K) and South Asian summary source (N=10K). When the genetic correlation was 1, the multi-source prediction methods cannot improve prediction accuracy in comparison with the single-source prediction methods using European summary statistics, because Europeans shared the same true effect sizes as South Asians and had ten times sample size. In the scenario where the genetic correlation was less than 1, multi-source prediction methods improved prediction accuracy over single-source prediction methods, reflecting the increase in source sample size. Overall, while lsum-multi outperformed PT-multi and PRS-CSx in most cases, MTL-PRS-lsum further improved cross-population prediction accuracy comparing lsum-multi across all simulation settings (Figure 4.3 & 4.4). For example, when genetic correlation was 0.4, TL-PRS-lsum improved a 6.49% average prediction accuracy compared to lsum-multi when the causal variants were 0.1%.

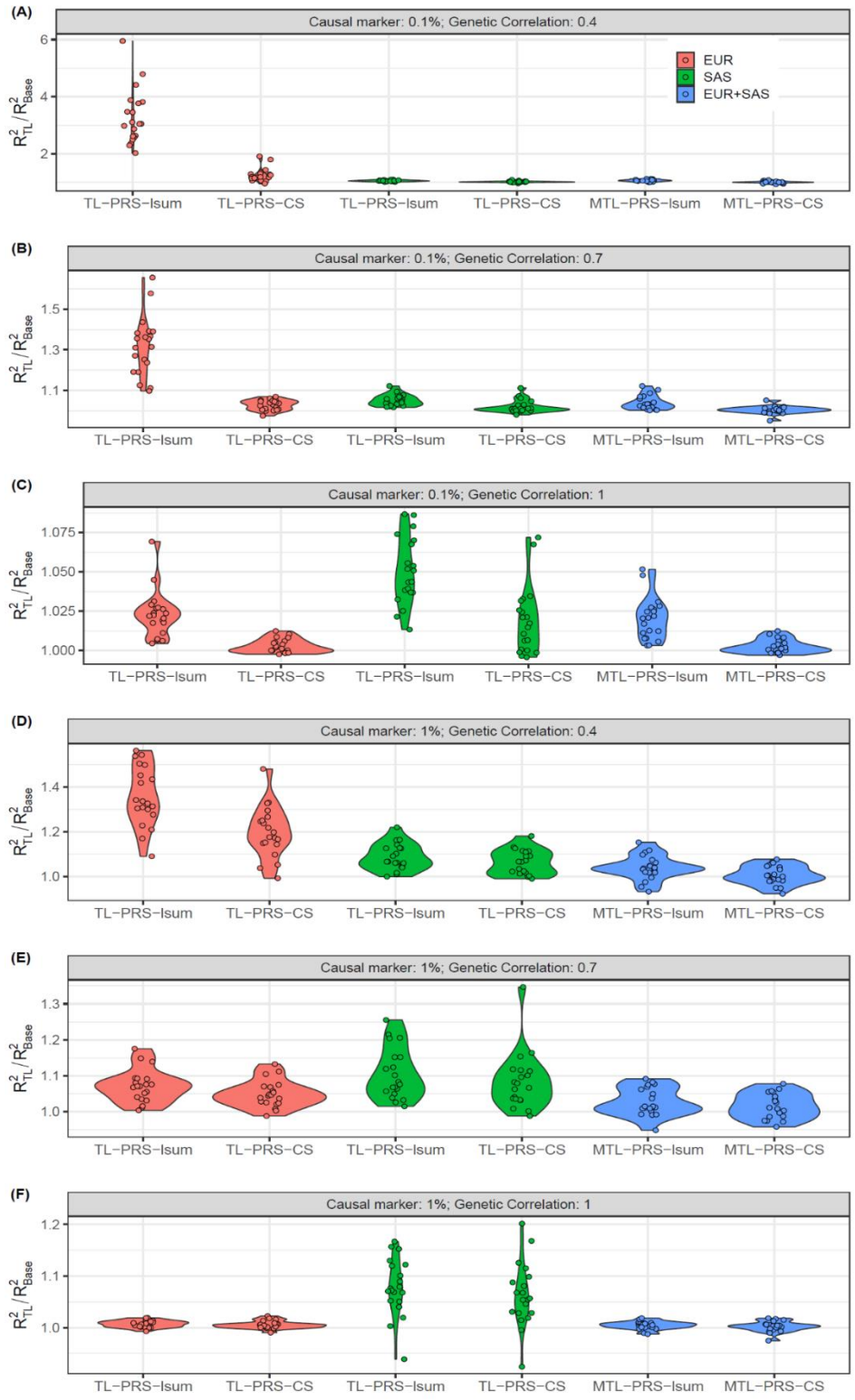


Figure 4.4. Relative prediction accuracy of single-source and multi-source polygenic prediction methods using transfer learning, with respect to the base models without transfer learning across 20 replicates in the simulation.

4.3.3 Prediction performance for South Asian, African, and non-British White samples in UK Biobank

After excluding related individuals, the target sample size of SAS, AFR, and NBW were 10,285, 8,168 and 35,567, respectively. We randomly split them into training dataset (for model fitting), validation dataset (for hyper-parameter tuning) and testing dataset (for the evaluation of predictive performance) (Supplementary Table S4.1). We applied single-source prediction methods to the UKBB or BBJ GWAS summary results, and used multi-source prediction methods to combine the UKBB and BBJ GWAS results.

Table 4.1 shows the prediction accuracy of different PRS construction methods in analyses of LDL in the African cohort of UK Biobank. When using UKBB GWAS results, the prediction R^2 of TL-PRS-lsum (0.044) and TL-PRS-CS (0.025) was higher than lsum (0.026) and PRS-CS (0.023). In addition, when using BBJ GWAS results, the prediction R^2 of TL-PRS-lsum (0.075) and TL-PRS-CS (0.034) was higher than lsum (0.048) and PRS-CS (0.030), demonstrating higher prediction accuracy in TL-PRS models. When combining UKBB and BBJ GWAS results, both lsum-multi (0.049) and PRS-CSx (0.043) outperformed PT-multi (0.030), as expected. At the same time, MTL-PRS-lsum (0.073) and MTL-PRS-CS (0.044) reached the best prediction accuracy. These consistent conclusions were reached when using the criteria of beta coefficients of normalized PRS or the mean difference between top 10% and bottom 10% PRS. The detailed results of other traits in the SAS, AFR and NBW populations can be found in Supplementary Table S4.2, S4.3 and S4.4, respectively.

Table 4.1. Prediction accuracy of different single-source and multi-source polygenic prediction methods in analyses of LDL in the African cohort of UK Biobank.

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction R^2 of PRS	Mean difference between top 10% and bottom 10% PRS
PT (UKBB)	0.095	-	0.012	0.388
lsum (UKBB)	0.139	-	0.026	0.543
TL-PRS-lsum (UKBB)	0.179	-	0.044	0.684
PRS-CS (UKBB)	0.136	-	0.023	0.448
TL-PRS-CS (UKBB)	0.345	-	0.025	0.624
PT (BBJ)	-	0.145	0.028	0.474
lsum (BBJ)	-	0.187	0.048	0.585
TL-lsum (BBJ)	-	0.235	0.075	0.913
PRS-CS (BBJ)	-	0.148	0.03	0.484
TL-PRS-CS (BBJ)	-	0.19	0.034	0.651
PT-multi	0.081	0.099	0.03	0.565
lsum-multi	0.074	0.137	0.049	0.696
MTL-PRS-lsum	0.061	0.183	0.073	0.891
PRSCSx	0.087	0.131	0.043	0.662
MTL-PRS-CS	0.261	0.14	0.044	0.776

Overall, consistent with the simulation results, TL-PRS-lsum and TL-PRS-CS outperformed lsum and PRS-CS in most traits from SAS and AFR (Figure 4.5). When the target population was SAS, TL-PRS-lsum attained 17% and 5% average relative improvement in prediction accuracy using

BBJ and UKBB GWAS results compared to lsum; the relative improvement of TL-PRS-CS over PRS-CS was on average 47% and 13% respectively. When the target population was AFR, TL-PRS-lsum attained 21% and 26% relative improvement of prediction accuracy in BBJ and UKBB GWAS results compared to lsum; TL-PRS-CS improved prediction accuracy by 40% and 9% compared to PRS-CS. When combining BBJ and UKBB GWAS results, MTL-PRS-lsum and MTL-PRS-CS had higher prediction performance than lsum-multi and PRS-CSx (Figure 4.5 & Figure S4.1). However, since non-British white samples were genetically similar to British white samples, TL-PRS-lsum, TL-PRS-CS, MTL-PRS-lsum, and MTL-PRS-CS all had nearly the identical performance to lsum, PRS-CS, lsum-multi and PRS-CSx (Figure 4.5). In general, the performance of TL-PRS depends on many factors, such as target populations, traits and baseline methods.

Figure S4.2 shows the cumulant event plot using the samples in the top 10% PRS across three populations. Across all situations, TL-PRS methods were found to have a similar or higher cumulant event curve than the baseline method. For example, in the analysis of CAD in the AFR cohort, when the age was up to 80, the cumulant prevalence of the samples with the top 10% PRS constructed by MTL-PRS-lsum was 0.26 while the prevalence in the samples with the top 10% PRS using lsum-multi was 0.12, suggesting that the TL method can be useful to help current PRS models improve the prediction of individualized disease risk and trajectories in certain diseases.

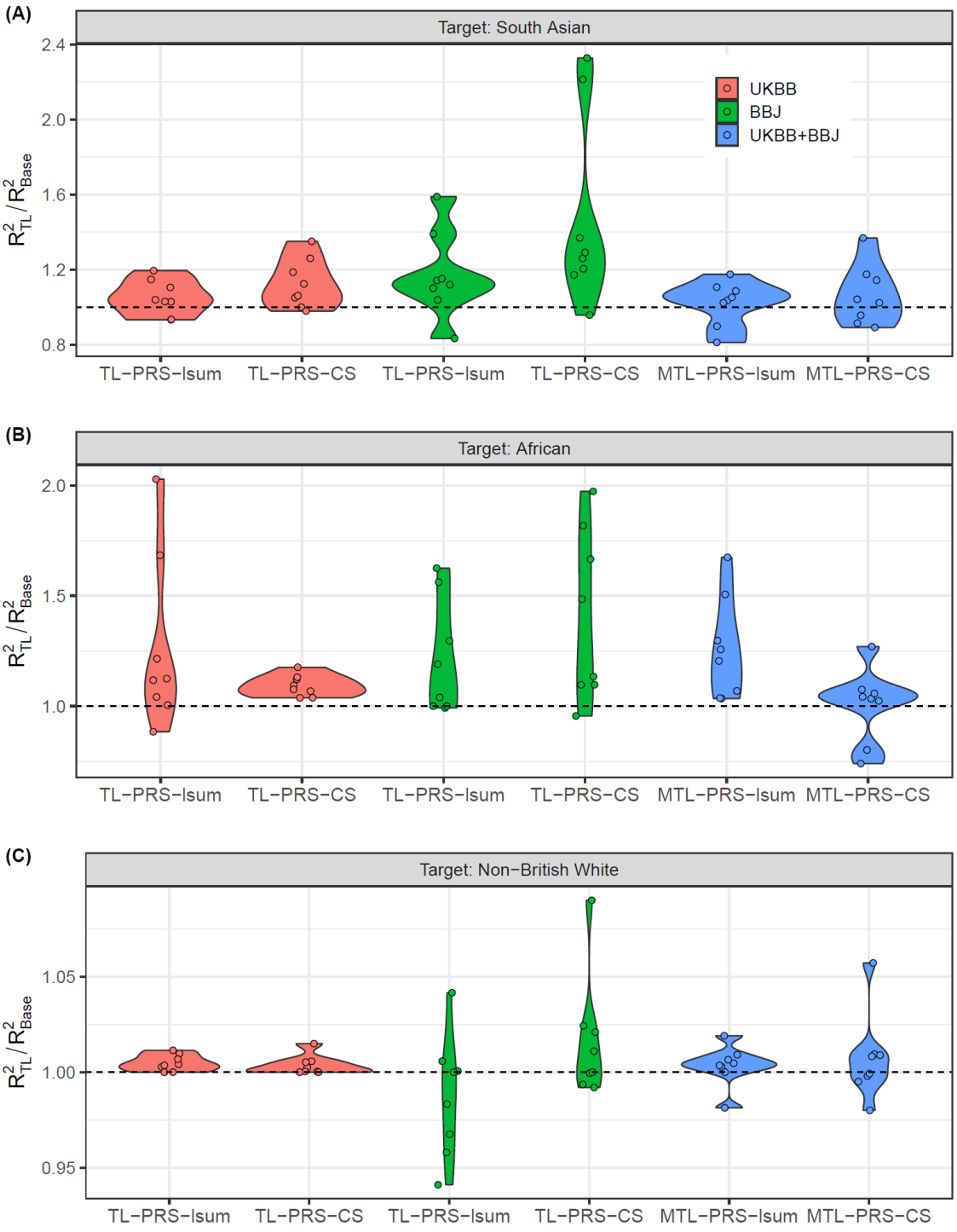


Figure 4.5. Relative prediction accuracy of single-source and multi-source polygenic prediction methods using transfer learning, with respect to the base models without transfer learning across 8 traits in South Asian, African and Non-British White.

4.4 Discussion

We have presented the TL-PRS method, which can transfer the existing biased PRS model from other populations to the target population. We have shown, through simulation studies, that TL-PRS-lsum and TL-PRS-CS robustly improves cross-population prediction over PT, lsum and PRS-CS across traits with varying genetic architectures, genetic correlations between target populations and samples used for calculating summary statistics. Using both quantitative and dichotomous traits from SAS, AFR, NBW populations in UK Biobank, we have demonstrated the TL-PRS can leverage large-scale European GWAS to boost the accuracy of polygenic prediction in non-European populations, for which ancestry-matched GWAS results may be orders of magnitude smaller in sample size.

Overall, the performance of TL-PRS depends on many factors, such as target populations, traits and baseline methods. When genetic correlations between target populations and samples used for calculating summary statistics are large, the baseline methods are less biased, thereby sufficient for prediction and TL-PRS might not further improve the prediction performance. When genetic correlations are small, TL-PRS can be applied on the target data to help transfer biased initial values to less biased effect sizes. TL-PRS appears to be most effective when genetic correlations are small. For example, the relative improvement of TL-PRS in the AFR and SAS cohorts was much larger than the NBW cohort, because AFR and SAS were genetically more different from European populations than NBW. Moreover, the approach could be further extended to admixed populations with simple modifications. Future work is needed to better evaluate the performance in admixed populations.

In addition, TL-PRS requires the individual-level data of the training dataset. When the individual-level data are not available, TL-PRS can still be applied given GWAS summary results of the targe samples. For example, $G_j Y$ can be estimated by the GWAS summary results and $G_j^T G$ can be estimated by the public available genotype data, such as 1000 Genome Project. Moreover, when the validation dataset is not available, the pseudo-validation(Mak et al., 2017) could further be applied for tuning the hyperparameters.

Despite these advantages, our work is subject to limitations and leaves several questions open for future exploration. First, although we have demonstrated large relative improvements in prediction accuracy, absolute prediction accuracies are currently not sufficiently high to achieve clinical utility for most traits (Chatterjee et al., 2013; Dudbridge, 2013); our simulations suggest that multi-ethnic polygenic risk scores will continue to produce improvements when more GWAS results are available, and the sample sizes are larger. Second, when combing two summary sources, the improvement of our MTL-PRS over the existing best PRS methods (PT-multi, lsum-multi and PRS-CSx) is limited. More research work is needed to combine more than one summary source. For example, the prediction performance might be further improved if cross-validation is used to determine the weights of the linear combination. The heritability of the traits, which differs across populations due to environmental factors, such as health behaviors and socioeconomic factors, can also be used to tune the weights of linear combination. Additionally, we did not incorporate data from the X chromosome, which is likely to harbor additional heritability that could improve prediction in some traits(Tukiainen et al., 2014). Finally, we restricted our analyses on common variants, but we may wish to incorporate the effects of rare variants in the future work. While extending present research to acquire more diverse genomes with sample sizes equivalent to European samples is the optimal, in the meantime, all existing available information should be

efficiently used to improve prediction across ancestries. TL-PRS increases the usefulness of PRS and reduces potential health inequities.

WEB RESOURCES

UK Biobank: https://www.ukbiobank.ac.uk/

BBJ summary statistics: http://jenger.riken.jp/en/result

1000 Genome Project: https://www.internationalgenome.org/

KING software: https://www.kingrelatedness.com/manual.shtml

Lassosum: https://github.com/tshmak/lassosum

PRS-CS: https://github.com/getian107/PRScs

PRS-CSx: https://github.com/getian107/PRScsx

TL-PRS: https://github.com/ZhangchenZhao/PRS_TransferLearning

4.5 Supplementary Materials

4.5.1 Supplementary Figures

Figure S4.1 Relative prediction accuracy of single-source and multi-source polygenic prediction methods with respect to lsum trained using UKBB GWAS across 8 traits in the cohorts South Asian, African and Non-British White. Each point shows the relative prediction R^2 of a trait.

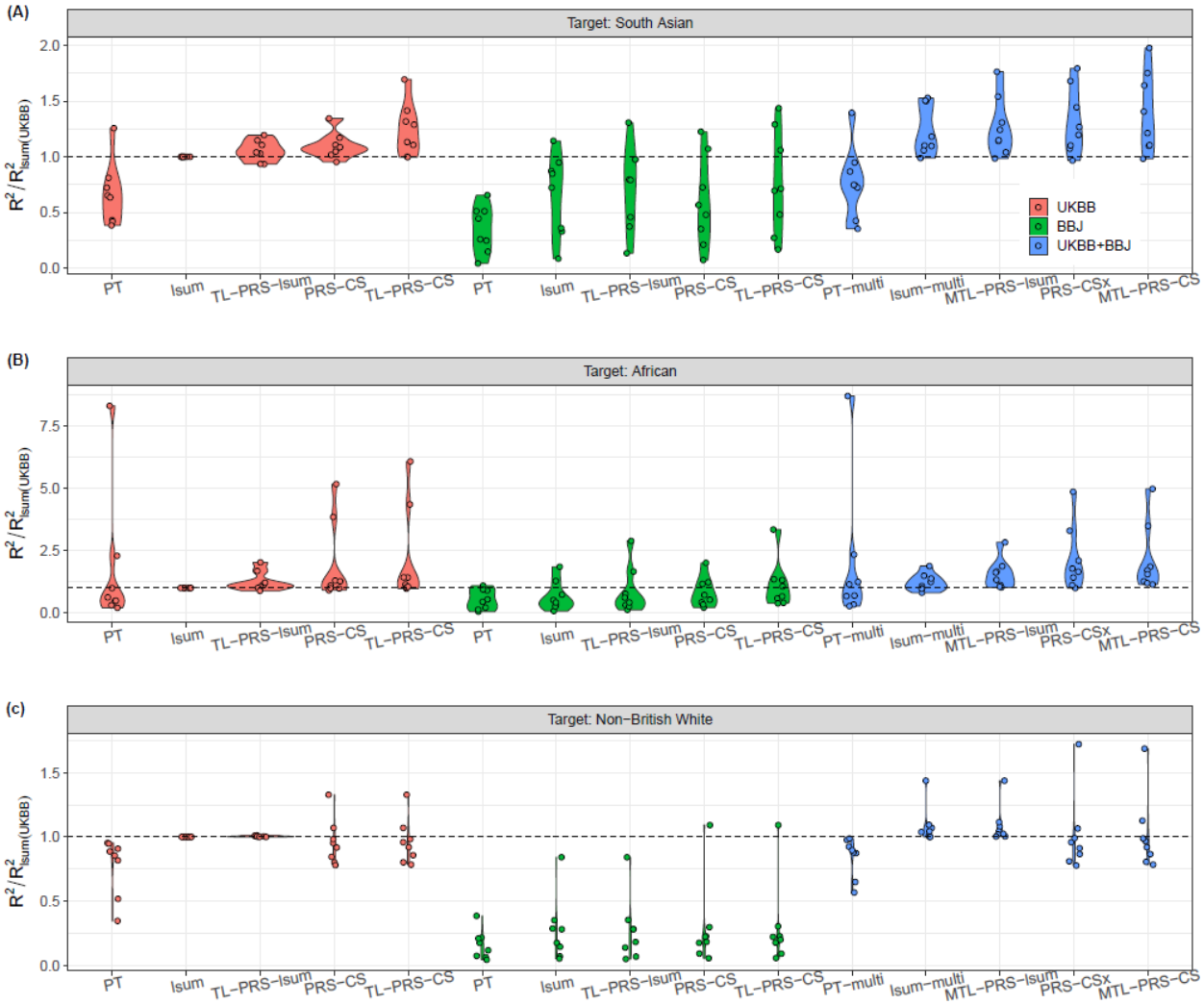
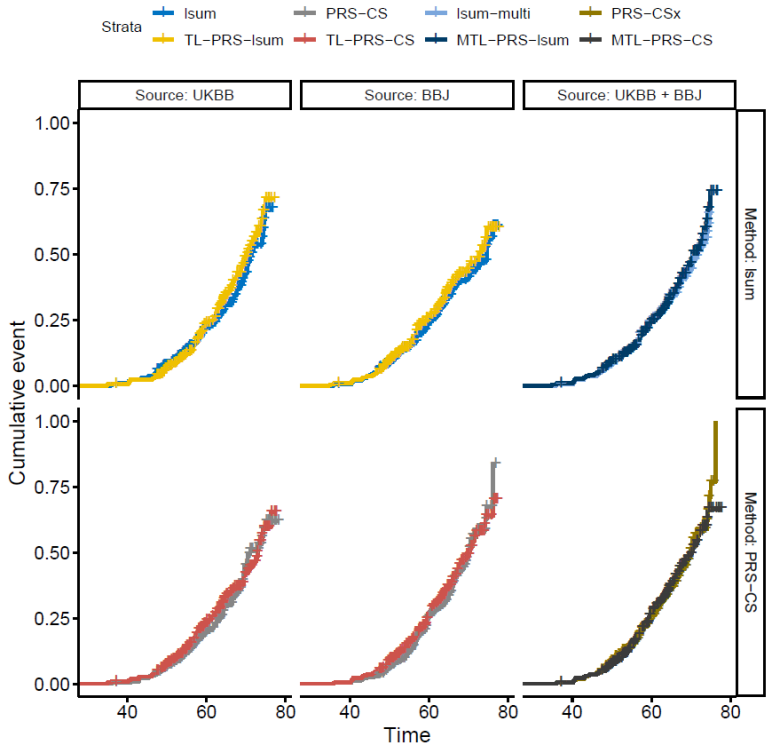
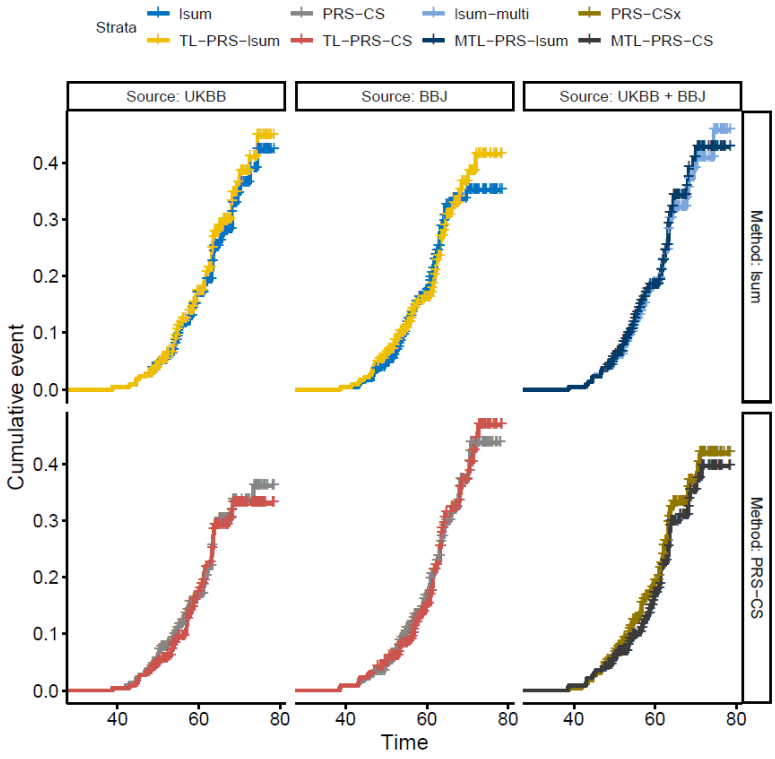


Figure S4.2. Cumulant event plot in terms of the top 10% PRS constructed by transfer learning methods and their base methods.

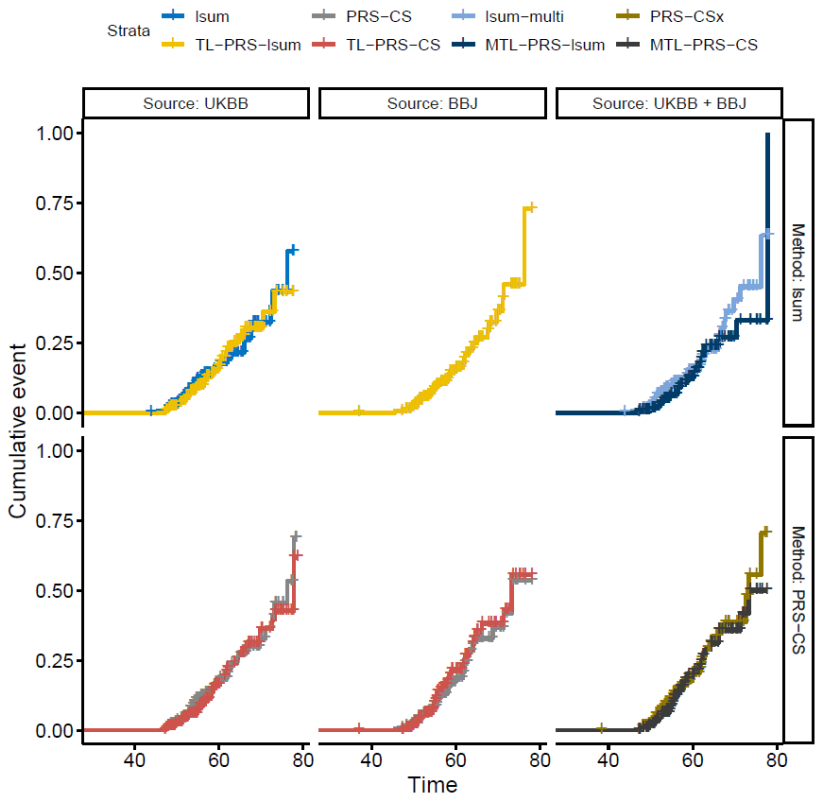
(a) South Asian, Type 2 diabetes, Case: Control=419:2211



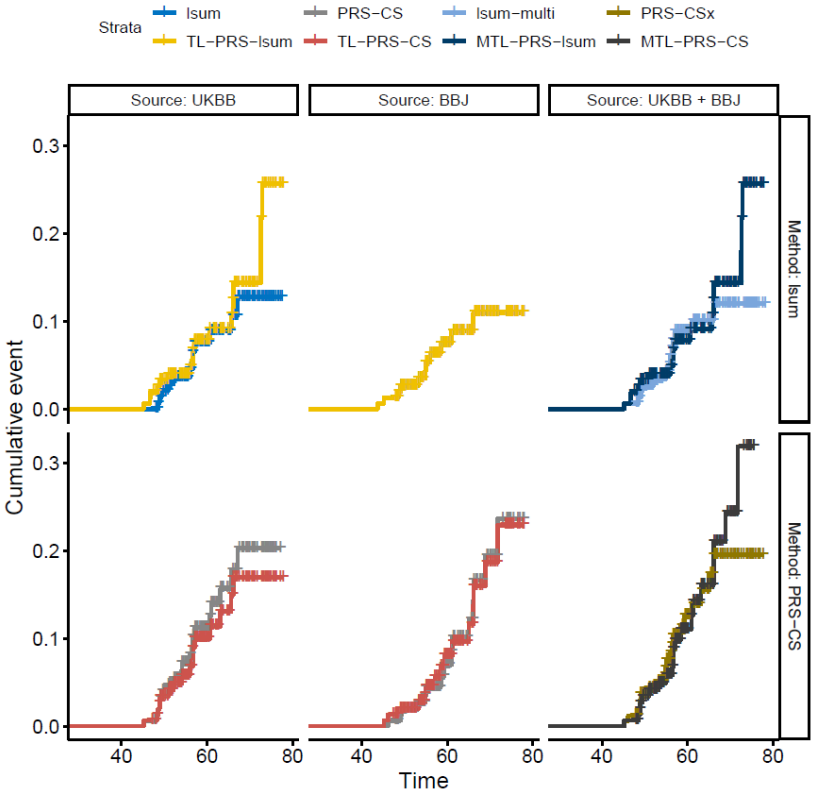
(b) South Asian, Coronary artery disease, Case: Control=362:2270



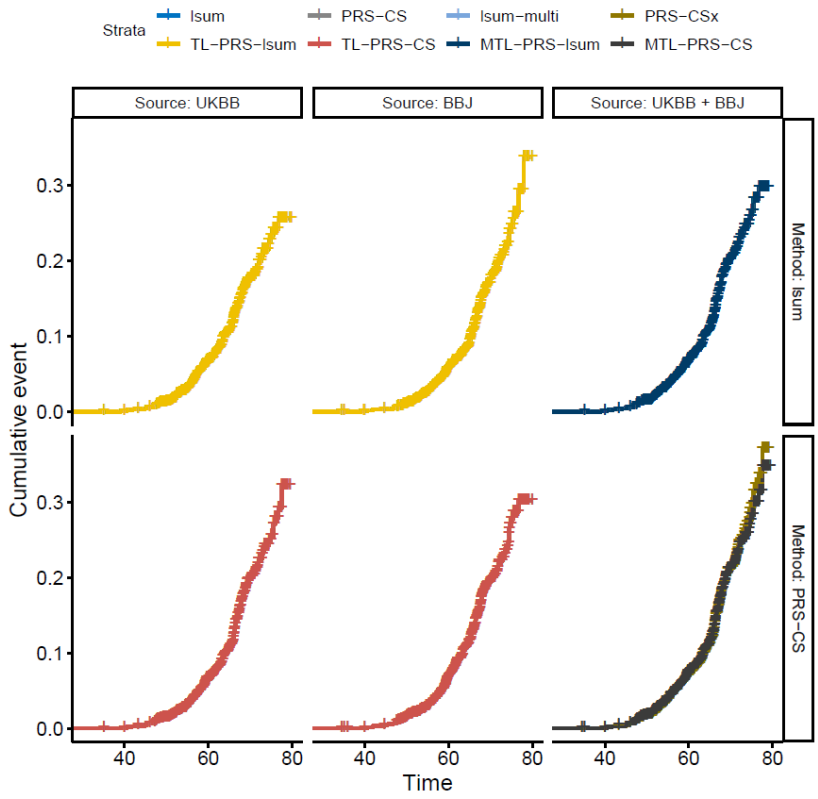
(c) African, Type 2 diabetes, Case: Control=177:1812



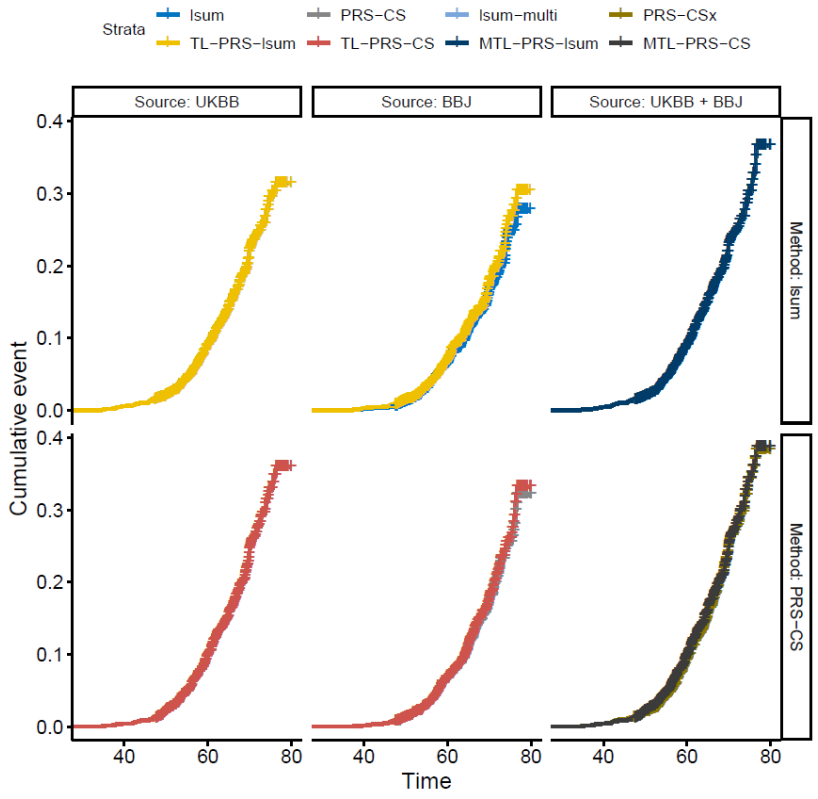
(d) African, Coronary artery disease, Case: Control=94:1902



(e) Non-British White, Type 2 diabetes, Case: Control=771:16737



(f) Non-British White, Coronary artery disease, Case: Control=1181:16357



4.5.2 Supplementary Tables

Table S4.1. List of data sets used in simulations and analyses of real phenotypes

Target Population	Trait	Total sample size	Training sample size	Validation sample size	Testing sample size
South Asian (SAS)	Simulation and real phenotypes	10,285	5,000	2,635	2,635
African (AFR)	real phenotypes	8,168	4,000	2,169	1,999
Non-British White (NBW)	real phenotypes	35,567	12,000	6,000	17,567

Table S4.2. Prediction accuracy of different PRS construction methods in analyses of eight traits in the South Asian cohort of UK Biobank

(a) HDL

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction R2 of PRS	Mean difference between top 10% and bottom 10% PRS
PT (UKBB)	0.085	-	0.07	0.301
lsum (UKBB)	0.094	-	0.086	0.3
TL-PRS-lsum (UKBB)	0.104	-	0.099	0.359
PRS-CS (UKBB)	0.102	-	0.101	0.336
TL-PRS-CS (UKBB)	0.111	-	0.114	0.365
PT (BBJ)	-	0.067	0.044	0.225
lsum (BBJ)	-	0.08	0.062	0.304
TL-lsum (BBJ)	-	0.083	0.069	0.293
PRS-CS (BBJ)	-	0.071	0.049	0.237
TL-PRS-CS (BBJ)	-	0.082	0.062	0.271
PT-multi	0.074	0.032	0.082	0.336
lsum-multi	0.076	0.041	0.102	0.364
MTL-PRS-lsum	0.087	0.037	0.113	0.39
PRSCSx	0.081	0.035	0.103	0.354
MTL-PRS-CS	0.09	0.039	0.121	0.368

(b) LDL

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction R2 of PRS	Mean difference between top 10% and bottom 10% PRS
PT (UKBB)	0.12	-	0.019	0.503
lsum (UKBB)	0.108	-	0.016	0.415
TL-PRS-lsum (UKBB)	0.12	-	0.019	0.447
PRS-CS (UKBB)	0.125	-	0.021	0.486
TL-PRS-CS (UKBB)	0.153	-	0.026	0.432
PT (BBJ)	-	0.077	0.008	0.206
lsum (BBJ)	-	0.114	0.018	0.445
TL-lsum (BBJ)	-	0.122	0.02	0.449
PRS-CS (BBJ)	-	0.074	0.007	0.252
TL-PRS-CS (BBJ)	-	0.115	0.016	0.414
PT-multi	0.094	0.062	0.022	0.459
lsum-multi	0.07	0.085	0.023	0.487
MTL-PRS-lsum	0.076	0.093	0.027	0.524
PRSCSx	0.107	0.046	0.022	0.505
MTL-PRS-CS	0.131	0.056	0.031	0.518

(c) BMI

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction R2 of PRS	Mean difference between top 10% and bottom 10% PRS
PT (UKBB)	0.953	-	0.044	2.939
lsum (UKBB)	1.175	-	0.067	3.598
TL-PRS-lsum (UKBB)	1.213	-	0.069	3.429
PRS-CS (UKBB)	1.142	-	0.064	3.763
TL-PRS-CS (UKBB)	1.208	-	0.067	3.591
PT (BBJ)	-	0.589	0.017	1.951
lsum (BBJ)	-	0.664	0.022	2.657
TL-lsum (BBJ)	-	0.785	0.031	2.955
PRS-CS (BBJ)	-	0.686	0.024	2.165
TL-PRS-CS (BBJ)	-	0.804	0.032	2.255
PT-multi	0.661	0.661	0.05	3.629
lsum-multi	0.945	0.509	0.071	4.117
MTL-PRS-lsum	0.987	0.531	0.077	4.198
PRSCSx	0.962	0.412	0.072	4.142
MTL-PRS-CS	0.954	0.514	0.074	3.698

(d) TG

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction R2 of PRS	Mean difference between top 10% and bottom 10% PRS
PT (UKBB)	0.256	-	0.053	0.787
lsum (UKBB)	0.301	-	0.074	1.058
TL-PRS-lsum (UKBB)	0.306	-	0.076	1.022
PRS-CS (UKBB)	0.317	-	0.082	1.032
TL-PRS-CS (UKBB)	0.326	-	0.082	1.022
PT (BBJ)	-	0.244	0.049	0.785
lsum (BBJ)	-	0.281	0.065	0.91
TL-lsum (BBJ)	-	0.298	0.072	0.929
PRS-CS (BBJ)	-	0.256	0.054	0.943
TL-PRS-CS (BBJ)	-	0.253	0.051	0.872
PT-multi	0.218	0.094	0.064	0.958
lsum-multi	0.268	0.067	0.082	1.124
MTL-PRS-lsum	0.272	0.068	0.084	1.084
PRSCSx	0.263	0.113	0.094	1.144
MTL-PRS-CS	0.282	0.094	0.09	1.041

(e) SBP

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction R2 of PRS	Mean difference between top 10% and bottom 10% PRS
PT (UKBB)	3.101	-	0.023	10.094
lsum (UKBB)	3.867	-	0.036	14.675
TL-PRS-lsum (UKBB)	3.782	-	0.034	15.388
PRS-CS (UKBB)	3.908	-	0.037	13.831
TL-PRS-CS (UKBB)	3.976	-	0.036	13.714
PT (BBJ)	-	1.479	0.005	5.576
lsum (BBJ)	-	2.328	0.013	8.706
TL-lsum (BBJ)	-	2.539	0.014	8.008
PRS-CS (BBJ)	-	1.778	0.008	8.38
TL-PRS-CS (BBJ)	-	2.461	0.01	8.176
PT-multi	2.926	1.254	0.026	12.106
lsum-multi	3.836	0.677	0.04	15.125
MTL-PRS-lsum	3.744	0.416	0.036	14.725
PRSCSx	3.643	0.911	0.04	14.612
MTL-PRS-CS	3.53	0.883	0.036	12.329

(f) DBP

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction R2 of PRS	Mean difference between top 10% and bottom 10% PRS
PT (UKBB)	1.309	-	0.014	5.153
lsum (UKBB)	2	-	0.033	7.435
TL-PRS-lsum (UKBB)	2.1	-	0.034	7.42
PRS-CS (UKBB)	2.063	-	0.035	6.973
TL-PRS-CS (UKBB)	2.285	-	0.037	7.753
PT (BBJ)	-	0.4	0.001	1.102
lsum (BBJ)	-	0.575	0.003	2.216
TL-lsum (BBJ)	-	0.735	0.004	2.706
PRS-CS (BBJ)	-	0.529	0.002	2.643
TL-PRS-CS (BBJ)	-	0.872	0.005	2.949
PT-multi	1.108	0.596	0.014	4.467
lsum-multi	1.801	0.6	0.033	7.2
MTL-PRS-lsum	1.872	0.624	0.034	6.663
PRSCSx	1.706	0.569	0.032	5.451
MTL-PRS-CS	2.088	0.368	0.036	7.494

(g) CAD

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction pseudo R2 of PRS	AUC of	Risk ratio between top 10% and bottom 10% PRS
PT (UKBB)	0.21	-	0.009	0.005	1.741
lsum (UKBB)	0.33	-	0.022	0.013	2.611
TL-PRS-lsum (UKBB)	0.319	-	0.02	0.012	2.136
PRS-CS (UKBB)	0.339	-	0.023	0.013	3.462
TL-PRS-CS (UKBB)	0.48	-	0.031	0.018	2.75
PT (BBJ)	-	0.163	0.005	0.002	1.273
lsum (BBJ)	-	0.321	0.02	0.01	2
TL-lsum (BBJ)	-	0.293	0.017	0.011	2.238
PRS-CS (BBJ)	-	0.34	0.023	0.012	4
TL-PRS-CS (BBJ)	-	0.407	0.028	0.016	2.286
PT-multi	0.033	0.189	0.008	0.003	1.171
lsum-multi	0.232	0.283	0.033	0.017	2.882
MTL-PRS-lsum	0.157	0.292	0.027	0.016	2.938
PRSCSx	0.347	0.187	0.039	0.022	4.077
MTL-PRS-CS	0.412	0.137	0.035	0.02	3.833

(h) Type II diabetes

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction pseudo R2 of PRS	AUC	Risk ratio between top 10% and bottom 10% PRS
PT (UKBB)	0.294	-	0.012	0.009	2.44
lsum (UKBB)	0.392	-	0.031	0.025	2.37
TL-PRS-lsum (UKBB)	0.46	-	0.035	0.026	3.65
PRS-CS (UKBB)	0.402	-	0.034	0.027	2.52
TL-PRS-CS (UKBB)	0.564	-	0.04	0.033	3.095
PT (BBJ)	-	0.249	0.014	0.01	1.867
lsum (BBJ)	-	0.352	0.027	0.023	2.739
TL-lsum (BBJ)	-	0.468	0.031	0.025	6.4
PRS-CS (BBJ)	-	0.421	0.038	0.027	2.8
TL-PRS-CS (BBJ)	-	0.514	0.045	0.031	3.789
PT-multi	0.244	0.244	0.023	0.018	2.583
lsum-multi	0.337	0.276	0.047	0.038	5.214
MTL-PRS-lsum	0.348	0.348	0.048	0.037	5.917
PRSCSx	0.319	0.261	0.053	0.039	3.714
MTL-PRS-CS	0.273	0.41	0.055	0.039	5.214

Table S4.3. Prediction accuracy of different PRS construction methods in analyses of eight traits in the African cohort of UK Biobank

(a) HDL

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction R2 of PRS	Mean difference between top 10% and bottom 10% PRS
PT (UKBB)	0.075	-	0.039	0.219
lsum (UKBB)	0.075	-	0.039	0.23
TL-PRS-lsum (UKBB)	0.085	-	0.044	0.204
PRS-CS (UKBB)	0.086	-	0.051	0.236
TL-PRS-CS (UKBB)	0.094	-	0.056	0.288
PT (BBJ)	-	0.07	0.035	0.259
lsum (BBJ)	-	0.083	0.05	0.316
TL-lsum (BBJ)	-	0.094	0.065	0.344
PRS-CS (BBJ)	-	0.082	0.048	0.328
TL-PRS-CS (BBJ)	-	0.089	0.053	0.311
PT-multi	0.057	0.038	0.049	0.268
lsum-multi	0.069	0.037	0.058	0.308
MTL-PRS-lsum	0.075	0.05	0.073	0.325
PRSCSx	0.071	0.047	0.07	0.312
MTL-PRS-CS	0.086	0.037	0.072	0.288

(b) LDL

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction R2 of PRS	Mean difference between top 10% and bottom 10% PRS
PT (UKBB)	0.095	-	0.012	0.388
lsum (UKBB)	0.139	-	0.026	0.543
TL-PRS-lsum (UKBB)	0.179	-	0.044	0.684
PRS-CS (UKBB)	0.136	-	0.023	0.448
TL-PRS-CS (UKBB)	0.345	-	0.025	0.624
PT (BBJ)	-	0.145	0.028	0.474
lsum (BBJ)	-	0.187	0.048	0.585
TL-lsum (BBJ)	-	0.235	0.075	0.913
PRS-CS (BBJ)	-	0.148	0.03	0.484
TL-PRS-CS (BBJ)	-	0.19	0.034	0.651
PT-multi	0.081	0.099	0.03	0.565
lsum-multi	0.074	0.137	0.049	0.696
MTL-PRS-lsum	0.061	0.183	0.073	0.891
PRSCSx	0.087	0.131	0.043	0.662
MTL-PRS-CS	0.261	0.14	0.044	0.776

(c) BMI

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction R2 of PRS	Mean difference between top 10% and bottom 10% PRS
PT (UKBB)	0.813	-	0.023	2.583
lsum (UKBB)	1.035	-	0.037	3.646
TL-PRS-lsum (UKBB)	1.179	-	0.038	3.477
PRS-CS (UKBB)	1.036	-	0.036	3.481
TL-PRS-CS (UKBB)	1.126	-	0.038	3.656
PT (BBJ)	-	0.476	0.008	1.242
lsum (BBJ)	-	0.581	0.012	2.375
TL-lsum (BBJ)	-	0.581	0.012	2.23
PRS-CS (BBJ)	-	0.748	0.019	2.631
TL-PRS-CS (BBJ)	-	0.786	0.021	2.732
PT-multi	0.653	0.435	0.025	2.682
lsum-multi	0.857	0.462	0.04	4.249
MTL-PRS-lsum	1.002	0.43	0.041	3.439
PRSCSx	0.927	0.232	0.037	3.158
MTL-PRS-CS	1.031	0.442	0.047	4.366

(d) TG

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction R2 of PRS	Mean difference between top 10% and bottom 10% PRS
PT (UKBB)	0.053	-	0.006	0.172
lsum (UKBB)	0.076	-	0.013	0.27
TL-PRS-lsum (UKBB)	0.085	-	0.015	0.377
PRS-CS (UKBB)	0.079	-	0.014	0.295
TL-PRS-CS (UKBB)	0.083	-	0.015	0.411
PT (BBJ)	-	0.051	0.006	0.207
lsum (BBJ)	-	0.065	0.01	0.182
TL-lsum (BBJ)	-	0.067	0.01	0.219
PRS-CS (BBJ)	-	0.049	0.006	0.172
TL-PRS-CS (BBJ)	-	0.05	0.005	0.096
PT-multi	0.049	0.026	0.009	0.228
lsum-multi	0.073	0.024	0.016	0.328
MTL-PRS-lsum	0.078	0.026	0.018	0.323
PRSCSx	0.076	0.025	0.019	0.323
MTL-PRS-CS	0.075	0.013	0.015	0.361

(e) SBP

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction R2 of PRS	Mean difference between top 10% and bottom 10% PRS
PT (UKBB)	1.163	-	0.003	4.451
lsum (UKBB)	2.062	-	0.011	6.287
TL-PRS-lsum (UKBB)	2.079	-	0.011	5.99
PRS-CS (UKBB)	2.052	-	0.01	9.282
TL-PRS-CS (UKBB)	2.226	-	0.011	9.174
PT (BBJ)	-	0.459	0.001	3.39
lsum (BBJ)	-	0.569	0.001	1.969
TL-lsum (BBJ)	-	0.729	0.001	3.508
PRS-CS (BBJ)	-	0.951	0.002	2.79
TL-PRS-CS (BBJ)	-	1.541	0.004	4.328
PT-multi	1.117	0.372	0.004	5.744
lsum-multi	1.907	0.477	0.01	5.379
MTL-PRS-lsum	1.951	0.488	0.011	5.446
PRSCSx	1.88	0.627	0.012	7.641
MTL-PRS-CS	2.07	0.887	0.013	8.533

(f) DBP

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction R2 of PRS	Mean difference between top 10% and bottom 10% PRS
PT (UKBB)	0.338	-	0.001	2.246
lsum (UKBB)	0.749	-	0.004	2.103
TL-PRS-lsum (UKBB)	0.844	-	0.005	1.959
PRS-CS (UKBB)	0.839	-	0.005	2.328
TL-PRS-CS (UKBB)	0.896	-	0.006	3.359
PT (BBJ)	-	0.268	0.001	0.877
lsum (BBJ)	-	0.535	0.002	1.821
TL-lsum (BBJ)	-	0.585	0.003	2.108
PRS-CS (BBJ)	-	0.431	0.001	1.641
TL-PRS-CS (BBJ)	-	0.642	0.003	3.621
PT-multi	0.376	0.094	0.001	1.503
lsum-multi	0.744	0.401	0.006	2.564
MTL-PRS-lsum	0.829	0.446	0.007	2.236
PRSCSx	0.942	0.314	0.009	3.59
MTL-PRS-CS	0.91	0.101	0.007	3.497

(g) CAD

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction pseudo R2 of PRS	AUC of	Risk ratio between top 10% and bottom 10% PRS
PT (UKBB)	0.448	-	0.028	0.03	3.167
lsum (UKBB)	0.154	-	0.003	0.004	3.667
TL-PRS-lsum (UKBB)	0.155	-	0.003	0.002	1.5
PRS-CS (UKBB)	0.3	-	0.013	0.019	2.571
TL-PRS-CS (UKBB)	0.332	-	0.014	0.017	2.286
PT (BBJ)	-	0.112	0.002	0.003	1.556
lsum (BBJ)	-	0.078	0.001	0.002	1.25
TL-lsum (BBJ)	-	0.078	0.001	0.002	1.25
PRS-CS (BBJ)	-	0.13	0.002	0.004	1.625
TL-PRS-CS (BBJ)	-	0.159	0.004	0.004	2.6
PT-multi	0.456	0.051	0.029	0.031	3
lsum-multi	0.071	0.106	0.003	0.004	1.5
MTL-PRS-lsum	0.141	0.076	0.003	0.003	1.571
PRSCSx	0.099	0.23	0.011	0.013	4.75
MTL-PRS-CS	0.164	0.201	0.012	0.013	3.167

(h) Type II diabetes

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction pseudo R2 of PRS	AUC	Risk ratio between top 10% and bottom 10% PRS
PT (UKBB)	0.278	-	0.012	0.005	1.118
lsum (UKBB)	0.172	-	0.005	0.005	1.235
TL-PRS-lsum (UKBB)	0.284	-	0.01	0.009	1.176
PRS-CS (UKBB)	0.4	-	0.026	0.018	2.5
TL-PRS-CS (UKBB)	0.781	-	0.031	0.022	1.389
PT (BBJ)	-	0.161	0.005	0.003	1.643
lsum (BBJ)	-	0.11	0.002	0.001	1.643
TL-lsum (BBJ)	-	0.11	0.002	0.001	1.643
PRS-CS (BBJ)	-	0.244	0.01	0.003	2.083
TL-PRS-CS (BBJ)	-	0.337	0.017	0.006	2.5
PT-multi	0.139	0.209	0.012	0.006	1.412
lsum-multi	0.075	0.14	0.005	0.002	1.2
MTL-PRS-lsum	0.138	0.169	0.008	0.004	1.556
PRSCSx	0.251	0.205	0.025	0.014	1.786
MTL-PRS-CS	0.219	0.329	0.025	0.013	1.647

Table S4.4. Prediction accuracy of different PRS construction methods in analyses of eight traits in the Non-British White cohort of UK Biobank

(a) HDL

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction R2 of PRS	Mean difference between top 10% and bottom 10% PRS
PT (UKBB)	0.154	-	0.153	0.537
lsum (UKBB)	0.157	-	0.16	0.559
TL-PRS-lsum (UKBB)	0.16	-	0.162	0.541
PRS-CS (UKBB)	0.154	-	0.153	0.54
TL-PRS-CS (UKBB)	0.156	-	0.154	0.521
PT (BBJ)	-	0.073	0.035	0.246
lsum (BBJ)	-	0.093	0.056	0.322
TL-lsum (BBJ)	-	0.093	0.057	0.322
PRS-CS (BBJ)	-	0.086	0.048	0.303
TL-PRS-CS (BBJ)	-	0.088	0.049	0.291
PT-multi	0.146	0.026	0.157	0.539
lsum-multi	0.14	0.047	0.172	0.57
MTL-PRS-lsum	0.143	0.048	0.173	0.564
PRSCSx	0.145	0.016	0.154	0.541
MTL-PRS-CS	0.148	0.016	0.155	0.525

(b) LDL

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction R2 of PRS	Mean difference between top 10% and bottom 10% PRS
PT (UKBB)	0.275	-	0.101	0.991
lsum (UKBB)	0.282	-	0.106	1.013
TL-PRS-lsum (UKBB)	0.285	-	0.107	1.003
PRS-CS (UKBB)	0.279	-	0.104	0.977
TL-PRS-CS (UKBB)	0.28	-	0.104	0.981
PT (BBJ)	-	0.118	0.019	0.426
lsum (BBJ)	-	0.151	0.03	0.538
TL-lsum (BBJ)	-	0.149	0.03	0.538
PRS-CS (BBJ)	-	0.118	0.019	0.42
TL-PRS-CS (BBJ)	-	0.12	0.019	0.423
PT-multi	0.259	0.065	0.105	1.009
lsum-multi	0.251	0.107	0.116	1.081
MTL-PRS-lsum	0.255	0.109	0.118	1.072
PRSCSx	0.259	0.046	0.105	0.988
MTL-PRS-CS	0.262	0.046	0.105	0.988

(c) BMI

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction R2 of PRS	Mean difference between top 10% and bottom 10% PRS
PT (UKBB)	1.897	-	0.149	6.795
lsum (UKBB)	2.058	-	0.174	7.467
TL-PRS-lsum (UKBB)	2.071	-	0.175	7.373
PRS-CS (UKBB)	1.893	-	0.147	6.653
TL-PRS-CS (UKBB)	1.927	-	0.149	6.584
PT (BBJ)	-	0.704	0.02	2.573
lsum (BBJ)	-	0.857	0.03	3.201
TL-lsum (BBJ)	-	0.878	0.032	3.171
PRS-CS (BBJ)	-	0.878	0.032	3.188
TL-PRS-CS (BBJ)	-	0.944	0.035	2.992
PT-multi	1.833	0.323	0.153	6.905
lsum-multi	1.992	0.221	0.176	7.4
MTL-PRS-lsum	2.003	0.223	0.178	7.336
PRSCSx	1.871	0.098	0.151	6.66
MTL-PRS-CS	1.849	0.205	0.151	6.623

(d) TG

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction R2 of PRS	Mean difference between top 10% and bottom 10% PRS
PT (UKBB)	0.325	-	0.102	1.16
lsum (UKBB)	0.359	-	0.125	1.259
TL-PRS-lsum (UKBB)	0.36	-	0.125	1.278
PRS-CS (UKBB)	0.346	-	0.114	1.2
TL-PRS-CS (UKBB)	0.346	-	0.115	1.193
PT (BBJ)	-	0.165	0.026	0.58
lsum (BBJ)	-	0.191	0.035	0.685
TL-lsum (BBJ)	-	0.192	0.035	0.707
PRS-CS (BBJ)	-	0.172	0.028	0.623
TL-PRS-CS (BBJ)	-	0.171	0.028	0.617
PT-multi	0.308	0.077	0.109	1.219
lsum-multi	0.329	0.082	0.129	1.301
MTL-PRS-lsum	0.33	0.083	0.13	1.287
PRSCSx	0.325	0.036	0.114	1.209
MTL-PRS-CS	0.328	0.036	0.115	1.198

(e) SBP

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction R2 of PRS	Mean difference between top 10% and bottom 10% PRS
PT (UKBB)	6.041	-	0.091	20.95
lsum (UKBB)	6.281	-	0.1	22.002
TL-PRS-lsum (UKBB)	6.292	-	0.101	22.208
PRS-CS (UKBB)	5.629	-	0.081	19.795
TL-PRS-CS (UKBB)	5.631	-	0.081	20.112
PT (BBJ)	-	1.569	0.006	5.6
lsum (BBJ)	-	1.645	0.007	5.746
TL-lsum (BBJ)	-	1.639	0.007	4.93
PRS-CS (BBJ)	-	1.879	0.009	6.204
TL-PRS-CS (BBJ)	-	1.893	0.009	6.217
PT-multi	5.97	0.663	0.093	21.456
lsum-multi	6.281	0	0.1	22.002
MTL-PRS-lsum	6.292	0	0.101	22.208
PRSCSx	5.54	0.292	0.081	19.4
MTL-PRS-CS	5.56	0.293	0.081	20.457

(f) DBP

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction R2 of PRS	Mean difference between top 10% and bottom 10% PRS
PT (UKBB)	3.53	-	0.106	12.916
lsum (UKBB)	3.75	-	0.12	13.761
TL-PRS-lsum (UKBB)	3.786	-	0.12	13.821
PRS-CS (UKBB)	3.315	-	0.094	11.892
TL-PRS-CS (UKBB)	3.375	-	0.094	12.107
PT (BBJ)	-	0.781	0.005	2.892
lsum (BBJ)	-	0.854	0.006	2.836
TL-lsum (BBJ)	-	0.85	0.006	3.02
PRS-CS (BBJ)	-	0.882	0.007	2.89
TL-PRS-CS (BBJ)	-	0.928	0.007	3.373
PT-multi	3.525	0.186	0.107	12.811
lsum-multi	3.75	0	0.12	13.761
MTL-PRS-lsum	3.786	0	0.12	13.821
PRSCSx	3.31	0	0.093	11.588
MTL-PRS-CS	3.375	0	0.094	12.107

(g) CAD

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction pseudo R2 of PRS	AUC of	Risk ratio between top 10% and bottom 10% PRS
PT (UKBB)	0.293	-	0.015	0.009	2.377
lsum (UKBB)	0.415	-	0.028	0.018	3.246
TL-PRS-lsum (UKBB)	0.415	-	0.028	0.018	3.246
PRS-CS (UKBB)	0.426	-	0.03	0.02	3.672
TL-PRS-CS (UKBB)	0.426	-	0.03	0.02	3.672
PT (BBJ)	-	0.114	0.002	0.001	1.516
lsum (BBJ)	-	0.158	0.004	0.002	1.722
TL-lsum (BBJ)	-	0.155	0.004	0.002	1.831
PRS-CS (BBJ)	-	0.197	0.006	0.003	1.772
TL-PRS-CS (BBJ)	-	0.196	0.006	0.003	1.813
PT-multi	0.278	0.119	0.016	0.01	2.247
lsum-multi	0.385	0.128	0.029	0.018	3.169
MTL-PRS-lsum	0.368	0.158	0.029	0.018	3.103
PRSCSx	0.356	0.152	0.03	0.02	3.338
MTL-PRS-CS	0.392	0.131	0.032	0.02	3.353

(h) Type II diabetes

Model	Beta of normalized UKBB PRS	Beta of normalized BBJ PRS	Prediction pseudo R2 of PRS	AUC	Risk ratio between top 10% and bottom 10% PRS
PT (UKBB)	0.264	-	0.012	0.018	3.848
lsum (UKBB)	0.5	-	0.034	0.043	6.417
TL-PRS-lsum (UKBB)	0.5	-	0.034	0.043	6.417
PRS-CS (UKBB)	0.579	-	0.045	0.052	6.538
TL-PRS-CS (UKBB)	0.579	-	0.045	0.052	6.538
PT (BBJ)	-	0.31	0.013	0.017	3.053
lsum (BBJ)	-	0.457	0.029	0.033	5.1
TL-lsum (BBJ)	-	0.457	0.029	0.033	5.1
PRS-CS (BBJ)	-	0.522	0.037	0.041	5.862
TL-PRS-CS (BBJ)	-	0.522	0.037	0.041	5.862
PT-multi	0.303	0.202	0.022	0.031	5.909
lsum-multi	0.484	0.261	0.049	0.058	8.6
MTL-PRS-lsum	0.484	0.261	0.049	0.058	8.6
PRSCSx	0.523	0.224	0.059	0.065	10.111
MTL-PRS-CS	0.531	0.228	0.058	0.063	10.111

CHAPTER V

Summary and Discussion

5.1 Summary

The growth of biobank datasets has presented us with enormous opportunities to gain insights into genetic architecture of complex traits and diseases. However, there are also a number of obstacles, both in terms of statistical technique and interpretation. In this dissertation, we addressed some of the most challenging problems of analyzing large-scale biobank datasets. To address the inflation of type I error rates caused by the unbalanced case-control ratios, we developed the robust region-based tests within independent samples. The robust methods were further extended to GLMM to adjust for relatedness among samples. Finally, we proposed the multi-ethnic PRS models using transfer learning to improve disease risk prediction in minor ancestries.

In Chapter II, we proposed the robust SKAT/SKAT-O type region-based tests to account for unbalanced case-control ratios, where the single-variant score statistic was calibrated based on SPA and ER. Through simulation studies, the proposed method was shown to provide well-calibrated p-values. The proposed method had similar computation time as the unadjusted approaches and was scalable for large sample data. In our application, the UK Biobank whole exome sequence data analysis of 45,596 unrelated European samples and 791 PheCode

phenotypes identified 10 rare variant associations with p-value < 10^{-7}, including the reported association between *JAK2* and myeloproliferative disease.

In Chapter III, we extended the robust method to related samples. A scalable generalized mixed-model region-based association test, SAIGE-GENE, was proposed that could handle large sample sizes and account for unbalanced case-control ratios for binary traits. This method utilized state-of-the-art optimization strategies to reduce computational and memory cost, and hence was applicable to exome-wide and genome-wide region-based analysis for hundreds of thousands of samples. Through the analysis of the HUNT study of 69,716 Norwegian samples and the UK Biobank data of 408,910 White British samples, SAIGE-GENE efficiently analyzed large sample data (N > 400,000) with type I error rates well controlled.

In Chapter IV, using summary statistics calculated by SAIGE-GENE, we proposed a novel multi-ethnic PRS motivated by transfer learning from machine learning literature. Our approach, TL-PRS, fine-tuned the potentially biased model trained with GWAS summary statistics from the majority ancestry to the target dataset of the minority ancestry. Through simulation studies, TL-PRS improved the performance of PRS with a wide range of genetic architectures and cross-population genetic correlations compared to the existing best PRS methods. In the application of 8,168 Africans and 10,285 South Asians of UK Biobank data, TL-PRS substantially improved the prediction accuracy of the six quantitative and two dichotomous traits.

In summary, these three chapters help us have a deep understanding of the associations between genotypes and phenotypes. Specifically, Chapter II and III aim to identify the true causal genes of a specific trait and Chapter IV quantifies the overall genetic effect using Chapter II & III summary results.

5.2 Extension and future work

The difficulties of high-dimensionality and computing scalability are extremely critical in this era of data explosion, and more efforts should be put on these issues in the future generations of methodological research. We addressed some of these problems in this dissertation, but the scope for future research is vast in this domain.

In all three chapters, we didn't consider potential selection bias when using UK biobank data. UK Biobank samples were recruited with response rate (5.5%) and it might be inaccurate to estimate the comorbidity of diseases (Swanson, 2012). But the disease prevalence is incorporated using the intercept in logistic regression and thus does not affect the detection of genetic effects or construction of PRS.

In Chapter III, when applying SAIGE-GENE on the latest 200K UKBB WES data, the inflation was observed in QQ plots when the case control ratio was extremely unbalanced, such as 1:500. When the sample size of UKBB WES data increased from 50K to 200K, more rare variants were identified in a single gene, which could cause the inflation of type I error rates. We are currently working on the new version of SAIGE-GENE by collapsing ultra-rare variants to control the inflation issue. In unbalanced case-control GWASs, scalable gene-based tests, and adjusting for within-study sample relatedness in meta-analysis methods, are also important and immediate future research directions. The similar robust adjustment could also be applied to other types of traits, such as time-to-event and categorical data.

In addition, interpreting GWAS results is a challenging process since the majority of phenotype-associated variants are non-coding and hence have no clear identifiable effect on protein functions as well as the phenotype. In the near future, integrating multi-omics data, such as proteome,

transcriptome, epigenome, metabolome, and microbiome, can further improve our understanding of the functional and mechanistic roles of different variants. The extension of GWAS, as well as its integration with other efforts, to understand the molecular function of the human genome, will be critically important in the study of gene coding and their contribution to complex traits. Further understanding on the genetic mechanism might help identify more disease biomarkers and drug targets.

In Chapter IV, although TL-PRS was shown to have large relative improvements in prediction accuracy using one summary source, the relative improvements of TL-PRS were limited compared to the best existing PRS method when combining more than one summary source. More investigations are needed to further improve the prediction accuracy of multi-ethnic PRS. For example, the performance might be improved if the cross-validation is used to determine the weights of the linear combination. Another possible approach is to incorporate multiple summary sources at the beginning of TL-PRS rather than linearly combining TL-PRS results. Additionally, the absolute prediction accuracies are currently not sufficiently high to achieve clinical utility for most traits (Chatterjee et al., 2013; Dudbridge, 2013). When more diverse genomes with sample sizes equivalent to European samples are available in the future, the multi-ethnic PRS will continue to improve the prediction performance.

Another future direction may include the exploration of the admixed populations. The gene-based models, such as robust methods and SAIGE-GENE, can be directly applied on admixed populations when the genotype sample sizes of admixed population were equivalent to European samples. In the intervening time, meta-analysis can help improve the detection power. In addition, TL-PRS is designed for the genetic risk prediction of minor populations and thus can be further

extended to admixed populations with simple modifications. Future work is needed to better evaluate the performance in admixed populations.

The problems and opportunities are not only confined to the ones described above. As we continue to generate ever huge volumes of data, new types of issues will emerge, necessitating new types of solutions. As a biostatistician, our objective is to keep working and avoid becoming overwhelmed by data, and this dissertation is a crucial step in that direction.

BIBLIOGRAPHY

Arking, D. E., Pulit, S. L., Crotti, L., van der Harst, P., Munroe, P. B., Koopmann, T. T., . . . Newton-Cheh, C. (2014). Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization. *Nat Genet*, *46*(8), 826-836. https://doi.org/10.1038/ng.3014

Asakai, R., Davie, E. W., & Chung, D. W. (1987). Organization of the gene for human factor XI. *Biochemistry*, *26*(23), 7221-7228.

Aubry, W., Lieberthal, R., Willis, A., Bagley, G., Willis III, S. M., & Layton, A. (2013). Budget impact model: epigenetic assay can help avoid unnecessary repeated prostate biopsies and reduce healthcare spending. *American health & drug benefits*, *6*(1), 15.

Avron, H., & Toledo, S. (2011). Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix %J J. ACM. *58*(2), 1-34. https://doi.org/10.1145/1944345.1944349

Bandesh, K., Prasad, G., Giri, A. K., Kauser, Y., Upadhyay, M., Indico, . . . Bharadwaj, D. (2019). Genome-wide association study of blood lipids in Indians confirms universality of established variants. *J Hum Genet*, *64*(6), 573-587. https://doi.org/10.1038/s10038-019-0591-7

Baxter, E. J., Scott, L. M., Campbell, P. J., East, C., Fouroucla s, N., Swanton, S., . . . Curtin, N. (2005). Acquired mutation of the tyrosine kinase JAK2 in human myeloproliferative disorders. *The Lancet*, *365*(9464), 1054-1061.

Bi, W., Zhao, Z., Dey, R., Fritsche, L. G., Mukherjee, B., & Lee, S. (2019). A Fast and Accurate Method for Genome-Wide Scale Phenome-Wide G × E Analysis and Its Application to UK Biobank. *The American Journal of Human Genetics*. https://doi.org/https://doi.org/10.1016/j.ajhg.2019.10.008

Biobank, U. (2019). *UK Biobank - Exome Data Release FAQs*. https://www.ukbiobank.ac.uk/wp-content/uploads/2019/08/UKB-50k-Exome-Sequencing-Data-Release-July-2019-FAQs.pdf

Breslow, N. E., & Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, *88*(421), 9-25. https://doi.org/10.2307/2290687

Bush, W. S., Oetjens, M. T., & Crawford, D. C. (2016). Unravelling the human genome–phenome relationship using phenome-wide association studies. *Nature Reviews Genetics*, *17*(3), 129.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., . . . O'Connell, J. (2017). Genome-wide genetic data on~ 500,000 UK Biobank participants. *BioRxiv*, 166298.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., . . . Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203-209. https://doi.org/10.1038/s41586-018-0579-z

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., . . . O'Connell, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203.

Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S. J., & Park, J.-H. (2013). Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nature genetics*, *45*(4), 400-405.

Chen, H., Huffman, J. E., Brody, J. A., Wang, C., Lee, S., Li, Z., . . . Bis, J. C. (2019). Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. *The American Journal of Human Genetics*, *104*(2), 260-274.

Chen, H., Huffman, J. E., Brody, J. A., Wang, C., Lee, S., Li, Z., . . . Lin, X. (2018). Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole genome sequencing studies. *bioRxiv*.

Chen, H., Huffman, J. E., Brody, J. A., Wang, C., Lee, S., Li, Z., . . . Lin, X. (2019). Efficient Variant Set Mixed Model Association Tests for Continuous and Binary Traits in Large-Scale Whole-Genome Sequencing Studies. *Am J Hum Genet*, *104*(2), 260-274. https://doi.org/10.1016/j.ajhg.2018.12.012

Chen, H., Wang, C., Conomos, M. P., Stilp, A. M., Li, Z., Sofer, T., . . . Lin, X. (2016). Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *Am J Hum Genet*, *98*(4), 653-666. https://doi.org/10.1016/j.ajhg.2016.02.012

Chen, H., Wang, C., Conomos, M. P., Stilp, A. M., Li, Z., Sofer, T., . . . Celedón, J. C. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics*, *98*(4), 653-666.

Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. *New England journal of medicine*, *372*(9), 793-795.

Consortium, I. H. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, *467*(7311), 52.

Daniels, H. E. (1954). Saddlepoint Approximations in Statistics. *Ann. Math. Statist.*, *25*(4), 631-650. https://doi.org/10.1214/aoms/1177728652

Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A. E., Kwong, A., . . . Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nat Genet*, *48*(10), 1284-1287. https://doi.org/10.1038/ng.3656

Davis, T. A. (2006). *Direct Methods for Sparse Linear Systems (Fundamentals of Algorithms 2)*. Society for Industrial and Applied Mathematics.

Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., Mosley, J. D., . . . Bowton, E. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature biotechnology*, *31*(12), 1102.

Dewey, F. E., Murray, M. F., Overton, J. D., Habegger, L., Leader, J. B., Fetterolf, S. N., . . . Gonzaga-Jauregui, C. (2016). Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science*, *354*(6319), aaf6814.

Dey, R., Schmidt, E. M., Abecasis, G. R., & Lee, S. (2017). A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. *Am J Hum Genet*, *101*(1), 37-49. https://doi.org/10.1016/j.ajhg.2017.05.014

Dey, R., Schmidt, E. M., Abecasis, G. R., & Lee, S. (2017). A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. *The American Journal of Human Genetics*, *101*(1), 37-49.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genet*, *9*(3), e1003348.

Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., . . . Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. *Nature communications*, *10*(1), 1-9.

Eijgelsheim, M., Newton-Cheh, C., Sotoodehnia, N., de Bakker, P. I., Muller, M., Morrison, A. C., . . . O'Donnell, C. J. (2010). Genome-wide association analysis identifies multiple loci related to resting heart rate. *Hum Mol Genet*, *19*(19), 3885-3894. https://doi.org/10.1093/hmg/ddq303

Eppinga, R. N., Hagemeijer, Y., Burgess, S., Hinds, D. A., Stefansson, K., Gudbjartsson, D. F., . . . van der Harst, P. (2016). Identification of genomic loci associated with resting heart rate and shared genetic predictors with all-cause mortality. *Nat Genet*, *48*(12), 1557-1563. https://doi.org/10.1038/ng.3708

Ewing, C. M., Ray, A. M., Lange, E. M., Zuhlke, K. A., Robbins, C. M., Tembe, W. D., . . . Wang, Y. (2012). Germline mutations in HOXB13 and prostate-cancer risk. *New England Journal of Medicine*, *366*(2), 141-149.

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., & Smoller, J. W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature communications*, *10*(1), 1-10.

Gilmour, A. R., Thompson, R., & Cullis, B. R. (1995). Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. *Biometrics*, *51*(4), 1440-1450. https://doi.org/10.2307/2533274

Hawkins, A. K., & O'Doherty, K. C. (2011). " Who owns your poop?": insights regarding the intersection of human microbiome research and the ELSI aspects of biobanking and related studies. *BMC Medical Genomics*, *4*(1), 72.

Holm, H., Gudbjartsson, D. F., Arnar, D. O., Thorleifsson, G., Thorgeirsson, G., Stefansdottir, H., . . . Stefansson, K. (2010). Several common variants modulate heart rate, PR interval and QRS duration. *Nat Genet*, *42*(2), 117-122. https://doi.org/10.1038/ng.511

Huang, H., Ruan, Y., Feng, Y.-C. A., Chen, C.-Y., Lam, M., Sawa, A., . . . Ge, T. (2021). Improving polygenic prediction in ancestrally diverse populations.

Hutchinson, M. F. (1990). A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics - Simulation and Computation*, *19*(2), 433-450. https://doi.org/10.1080/03610919008812866

Kaasschieter, E. F. (1988). Preconditioned conjugate gradients for solving singular systems. *Journal of Computational and Applied Mathematics*, *24*(1), 265-275. https://doi.org/https://doi.org/10.1016/0377-0427(88)90358-5

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., . . . Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*, *42*(4), 348-354. https://doi.org/10.1038/ng.548

Krokstad, S., Langhammer, A., Hveem, K., Holmen, T. L., Midthjell, K., Stene, T. R., . . . Holmen, J. (2013). Cohort Profile: the HUNT Study, Norway. *Int J Epidemiol*, *42*(4), 968-977. https://doi.org/10.1093/ije/dys095

Kuonen, D. (1999). Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika*, *4*(86), 7.

Langhammer, A., Krokstad, S., Romundstad, P., Heggland, J., & Holmen, J. (2012). The HUNT study: participation is associated with survival and depends on socioeconomic status, diseases and symptoms. *BMC medical research methodology*, *12*, 143-143. https://doi.org/10.1186/1471-2288-12-143

Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet*, *95*(1), 5-23. https://doi.org/10.1016/j.ajhg.2014.06.009

Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics*, *95*(1), 5-23.

Lee, S., Fuchsberger, C., Kim, S., & Scott, L. (2015). An efficient resampling method for calibrating single and gene-based rare variant association analysis in case–control studies. *Biostatistics*, *17*(1), 1-15.

Lee, S., Fuchsberger, C., Kim, S., & Scott, L. (2016). An efficient resampling method for calibrating single and gene-based rare variant association analysis in case-control studies. *Biostatistics*, *17*(1), 1-15. https://doi.org/10.1093/biostatistics/kxv033

Lee, S., Teslovich, T. M., Boehnke, M., & Lin, X. (2013). General framework for meta-analysis of rare variants in sequencing association studies. *Am J Hum Genet*, *93*(1), 42-53. https://doi.org/10.1016/j.ajhg.2013.05.010

Lee, S., Wu, M. C., & Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, *13*(4), 762-775.

Lee, S., Wu, M. C., & Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, *13*(4), 762-775. https://doi.org/10.1093/biostatistics/kxs014

Lee, S. H., & van der Werf, J. H. (2006). An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. *Genet Sel Evol*, *38*(1), 25-43. https://doi.org/10.1051/gse:2005025

Lewis, C. M., & Vassos, E. (2020). Polygenic risk scores: from research tools to clinical instruments. *Genome medicine*, *12*, 1-11.

Li, B., & Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, *83*(3), 311-321.

Liu, D. J., Peloso, G. M., Zhan, X., Holmen, O. L., Zawistowski, M., Feng, S., . . . Abecasis, G. R. (2014). Meta-analysis of gene-level tests for rare variant association. *Nat Genet*, *46*(2), 200-204. https://doi.org/10.1038/ng.2852

Loh, P. R., Tucker, G., Bulik-Sullivan, B. K., Vilhjalmsson, B. J., Finucane, H. K., Salem, R. M., . . . Price, A. L. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet*, *47*(3), 284-290. https://doi.org/10.1038/ng.3190

Ma, C., Blackwell, T., Boehnke, M., Scott, L. J., & Investigators, G. D. (2013). Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genetic epidemiology*, *37*(6), 539-550.

Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X., & Sham, P. C. (2017). Polygenic scores via penalized regression on summary statistics. *Genetic epidemiology*, *41*(6), 469-480.

Marouli, E., Graff, M., Medina-Gomez, C., Lo, K. S., Wood, A. R., Kjaer, T. R., . . . Lettre, G. (2017). Rare and low-frequency coding variants alter human adult height. *Nature*, *542*(7640), 186-190. https://doi.org/10.1038/nature21039

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., & Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics*, *51*(4), 584-591.

Morgenthaler, S., & Thilly, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *615*(1-2), 28-56.

Márquez-Luna, C., Loh, P. R., Consortium, S. A. T. D., Consortium, S. T. D., & Price, A. L. (2017). Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genetic epidemiology*, *41*(8), 811-823.

Natarajan, P., Peloso, G. M., Zekavat, S. M., Montasser, M., Ganna, A., Chaffin, M., . . . Group, N. T. L. W. (2018). Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nat Commun*, *9*(1), 3391. https://doi.org/10.1038/s41467-018-05747-8

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*(10), 1345-1359.

Pendergrass, S., & Ritchie, M. D. (2015). Phenome-wide association studies: leveraging comprehensive phenotypic and genotypic data for discovery. *Current genetic medicine reports*, *3*(2), 92-100.

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., & Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, *460*(7256), 748-752.

Regier, A. A., Farjoun, Y., Larson, D. E., Krasheninina, O., Kang, H. M., Howrigan, D. P., . . . Ames, D. C. (2018). Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nature communications*, *9*(1), 4038.

Replication, D. I. G., Meta-analysis, C., Asian Genetic Epidemiology Network Type 2 Diabetes, C., South Asian Type 2 Diabetes, C., Mexican American Type 2 Diabetes, C., Type 2 Diabetes Genetic Exploration by Nex-generation sequencing in muylti-Ethnic Samples, C., . . . Morris, A. P. (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet*, *46*(3), 234-244. https://doi.org/10.1038/ng.2897

Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J., & Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome research*, *15*(11), 1576-1583.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., . . . Landray, M. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, *12*(3), e1001779.

Sugrue, L. P., & Desikan, R. S. (2019). What are polygenic scores and why are they important? *Jama*, *321*(18), 1820-1821.
Sung, Y. J., Winkler, T. W., de Las Fuentes, L., Bentley, A. R., Brown, M. R., Kraja, A. T., . . . Chasman, D. I. (2018). A Large-Scale Multi-ancestry Genome-wide Study Accounting for Smoking Behavior Identifies Multiple Significant Loci for Blood Pressure. *Am J Hum Genet*, *102*(3), 375-400. https://doi.org/10.1016/j.ajhg.2018.01.015
Surendran, P., Drenos, F., Young, R., Warren, H., Cook, J. P., Manning, A. K., . . . Munroe, P. B. (2016). Trans-ancestry meta-analyses identify rare and common variants associated with blood pressure and hypertension. *Nat Genet*, *48*(10), 1151-1161. https://doi.org/10.1038/ng.3654
Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., van Duijn, C. M., & Aulchenko, Y. S. (2012). Rapid variance components-based method for whole-genome association analysis. *Nat Genet*, *44*(10), 1166-1170. https://doi.org/10.1038/ng.2410
Swanson, J. M. (2012). The UK Biobank and selection bias. *The Lancet*, *380*(9837), 110.
Swoap, S. J., Weinshenker, D., Palmiter, R. D., & Garber, G. (2004). Dbh(-/-) mice are hypotensive, have altered circadian rhythms, and have abnormal responses to dieting and stress. *Am J Physiol Regul Integr Comp Physiol*, *286*(1), R108-113. https://doi.org/10.1152/ajpregu.00405.2003
Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., . . . Abecasis, G. R. (2019). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *bioRxiv*.
Torkamani, A., Wineinger, N. E., & Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, *19*(9), 581-590.
Torrey, L., & Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques* (pp. 242-264). IGI global.
Tukiainen, T., Pirinen, M., Sarin, A.-P., Ladenvall, C., Kettunen, J., Lehtimäki, T., . . . Vlachopoulou, E. (2014). Chromosome X-wide association study identifies Loci for fasting insulin and height and evidence for incomplete dosage compensation. *PLoS Genet*, *10*(2), e1004127.
Turalba, A. V., & Chen, T. C. (2008). Clinical and genetic characteristics of primary juvenile-onset open-angle glaucoma (JOAG). *Semin Ophthalmol*, *23*(1), 19-25. https://doi.org/10.1080/08820530701745199
Van Hout, C. V., Tachmazidou, I., Backman, J. D., Hoffman, J. X., Yi, B., Pandey, A., . . . Banerjee, N. (2019). Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. *bioRxiv*, 572347.
Verhaar, M., Stroes, E., & Rabelink, T. (2002). Folates and cardiovascular disease. *Arteriosclerosis, thrombosis, and vascular biology*, *22*(1), 6-13.
Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., . . . Do, R. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics*, *97*(4), 576-592.
Vranka, J. A., Sakai, L. Y., & Bächinger, H. P. (2004). Prolyl 3-hydroxylase 1, enzyme characterization and identification of a novel family of enzymes. *Journal of Biological Chemistry*, *279*(22), 23615-23621.

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, *38*(16), e164-e164.

Wang, L., Choi, S., Lee, S., Park, T., & Won, S. (2016). Comparing family-based rare variant association tests for dichotomous phenotypes. BMC proceedings,

West, J., Ventura, D., & Warnick, S. (2007). Spring research presentation: A theoretical foundation for inductive transfer. *Brigham Young University, College of Physical and Mathematical Sciences*, *1*(08).

Willer, C. J., Sanna, S., Jackson, A. U., Scuteri, A., Bonnycastle, L. L., Clarke, R., . . . Abecasis, G. R. (2008). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet*, *40*(2), 161-169. https://doi.org/10.1038/ng.76

Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., . . . Global Lipids Genetics, C. (2013). Discovery and refinement of loci associated with lipid levels. *Nat Genet*, *45*(11), 1274-1283. https://doi.org/10.1038/ng.2797

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, *89*(1), 82-93.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*, *89*(1), 82-93. https://doi.org/10.1016/j.ajhg.2011.05.029

Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., & Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet*, *46*(2), 100-106. https://doi.org/10.1038/ng.2876

Zengini, E., Hatzikotoulas, K., Tachmazidou, I., Steinberg, J., Hartwig, F. P., Southam, L., . . . Gilly, A. (2018). Genome-wide analyses using UK Biobank data provide insights into the genetic architecture of osteoarthritis. *Nature genetics*, *50*(4), 549.

Zhang, X., Basile, A. O., Pendergrass, S. A., & Ritchie, M. D. (2019). Real world scenarios in rare variant association analysis: the impact of imbalance and sample size on the power in silico. *BMC bioinformatics*, *20*(1), 46.

Zhao, Z., Bi, W., Zhou, W., VandeHaar, P., Fritsche, L. G., & Lee, S. (2019). UK-Biobank Whole Exome Sequence Binary Phenome Analysis with Robust Region-based Rare Variant Test. *bioRxiv*.

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., . . . Gifford, A. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature genetics*, *50*(9), 1335.

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., . . . Lee, S. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet*, *50*(9), 1335-1341. https://doi.org/10.1038/s41588-018-0184-y

Zhou, W., Nielsen, J. B., Fritsche, L. G., LeFaive, J., Taliun, S. A. G., Bi, W., . . . Hveem, K. (2019). Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *BioRxiv*, 583278.